DOCTEUR EN SCIENCES

Enseignement de l'inférence statistique

analyse en termes de transposition didactique et mise en place d'une ingénierie didactique

Bihin, Benoit

Award date:
2021

Awarding institution:
Universite de Namur

UNIVERSITÉ DE NAMUR

Enseignement de l'inférence statistique : analyse en termes de transposition didactique et mise en place d'une ingénierie didactique

Benoit Bihin

Dissertation en vue de l'obtention du grade de docteur en Sciences

Membres du jury :

M. Romainville, UNamur (Président)

V. Henry, UNamur

J. Plumat, UNamur

F. Farnir, ULiège

J.-C. Régnier, Lyon 2

B. Falissard, CESP/INSERM

E. Depiereux, UNamur (Co-promoteur)

A. Vervoort, UNamur (Co-promoteur)

Juin 2021

Résumé

L'inférence statistique est le champ de la statistique comprenant les outils mettant en relation des observations empiriques et des hypothèses statistiques telles que "l'efficacité de ce vaccin est de 80 %" ou "la prise de placebo diminue de moitié la douleur lors de crises de migraine". Parmi ces outils, on trouve : le test de significativité, mesurant à quel point des observations corroborent une hypothèse ; le test d'hypothèses, utilisant les observations pour opérer un choix parmi plusieurs hypothèses concurrentes ou encore l'intervalle de confiance, définissant les hypothèses statistiques compatibles avec une série d'observations. Situés à l'interface entre les observations et les hypothèses des chercheurs, ces outils statistiques occupent une place centrale dans l'analyse des données expérimentales et dans la construction des savoirs scientifiques. Ils sont, malheureusement, souvent mal utilisés par les chercheurs et difficiles à enseigner aux étudiants, notamment ceux des filières biomédicales à l'Université de Namur.

Dans ce travail, nous cherchons à améliorer l'enseignement de l'inférence statistique.

L'analyse du savoir enseigné localement révèle que le test d'hypothèses enseigné diffère sensiblement du test d'hypothèses initialement décrit.

L'analyse des pratiques professionnelles suggère que le test d'hypothèses enseigné s'inspire d'une pratique courante en recherche biomédicale, le *Null Hypothesis Significance Testing* (NHST). Celui-ci est pourtant très critiqué et, notamment, considéré en partie responsable de la crise de reproductibilité en sciences. L'examen d'une telle pratique montre qu'elle fait très peu intervenir le discernement du chercheur dans la confrontation entre les observations et l'hypothèse de recherche. Ceci la rend plus objective, plus facile à appliquer, à justifier mais aussi, à enseigner. Toutefois, en éludant l'énoncé préalable d'une hypothèse de recherche précise ainsi que la délicate question de l'interprétation de l'ampleur des effets, le NHST met en œuvre une version altérée de démarche scientifique.

Améliorer l'enseignement de l'inférence statistique aux filières biomédicales nécessiterait, dès lors, de le recentrer autour d'une véritable démarche de recherche. L'analyse d'une expérience d'enseignement de l'inférence statistique permet de soulever plusieurs pistes pour y parvenir.

Remerciements

Lorsqu'en 2011 j'ai débuté mon mandat d'assistant-doctorant, je voyais la thèse comme un travail dont le mérite était essentiellement individuel.

Les dix années qui suivirent m'ont prouvé que j'avais tort.

Bien sûr qu'une thèse implique un important travail personnel mais jamais je n'aurais été en mesure de le mener à son terme sans le concours de mentors, de collègues et de mes proches qui ont contribué à établir le cadre professionnel et privé rendant possible la rédaction d'une thèse. Je voudrais ici les remercier.

Merci à mes promoteurs, Eric et Arnaud, pour avoir m'avoir accordé leur confiance et m'avoir soutenu dans les moments cruciaux. La liberté qu'ils m'ont laissée dans la conduite de cette recherche s'est révélée très motivante et, au final, très formatrice.

Merci aux membres de mon comité d'accompagnement, Valérie Henry et Jim Plumat, pour le temps qu'ils m'ont consacré, pour s'être penché sur mon travail avec un recul critique, m'aidant ainsi à rendre mon travail plus pertinent et mon propos plus précis.

Merci également à Jérémy Dehon et à Marc Romainville pour leurs conseils avisés au moment de finaliser la structure de la thèse.

Merci au membres extérieurs de ce jury de thèse, Bruno Falissard, Frédéric Farnir et Jean-Claude Régnier, d'avoir accepté de faire partie de ce jury et pour leurs remarques constructives et encourageantes.

Puisque la fin de la thèse rime également avec mon départ de l'unité de recherche en didactique de la biologie (URDB), je profite de l'occasion pour remercier mes collègues passés et présents de l'unité avec qui j'ai pris beaucoup de plaisir à travailler ces dix dernières années : Bénédicte, Véronique, Sophie, Anne-Cécile, Grégoire, Mélanie, Marie-Laurence, Jean-Pierre, Etienne, Françoise, Marianne, Guitou. Un merci particulier aux assistantes et assistants qui ont accepté (ou été contraints) de s'adapter aux travaux pratiques que je proposais et à propos desquels la seule chose qui s'est avérée stable d'une année à l'autre c'était le changement.

Je salue également mes collègues de Mont-Godinne qui, depuis 2014, m'ont gentiment accueillis dans l'unité de support scientifique (USS) : Isabelle, Marie-Paule, Katty, Annouck, Michèle, Virginie, Ingrid, Thuyen, Cécile, Floriane, Fabienne, Axel, Christian, Marie-Bernadette et une pensée particulière pour Laurence. Merci à Maxime pour sa bonne humeur et la motivation dont il fait part. Sa grande implication dans la gestion des analyses statistiques ces derniers mois m'a été d'une grande aide pour finaliser la rédaction de la thèse. Merci à Jacques Jamart de m'avoir initié au métier de biostatisticien. Son expérience et sa disponibilité m'ont permis de bien démarrer dans le monde passionnant mais complexe de la recherche biomédicale. J'espère qu'il ne regrette pas de m'avoir transmis une partie de son savoir et qu'il trouvera dans l'aboutissement de cette thèse une réponse à l'une de ses questions favorites : "Et ta thèse, elle avance ?".

Un merci tout particulier à Madeleine et Isabelle qui ont relu$\not$ et corrigé$\not$, les nombreuses versions qui ont précédé le présent texte.

Merci à Jérémie et Sophie, à mes amis et cousins pour les bons moments passés et à venir. Merci à mes parents et à mes beaux-parents pour leur soutien de tous les instants. Merci à Élise sans qui je ne m'imaginais pas finir cette thèse.

Je dédie cette thèse à mes grands-parents ainsi qu'à mes enfants.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Pourquoi s'intéresser à l'enseignement de l'inférence statistique dans le domaine biomédical ?

A notre échelle, nous constatons, d'année en année, que les étudiants butent systématiquement sur les notions liées à l'inférence statistique. Ce problème s'observe aussi chez les chercheurs avec qui nous collaborons, nombreux à ne pas maitriser l'interprétation d'une notion aussi familière dans la littérature scientifique que la P-valeur, par exemple.

Ce constat local n'est certes pas suffisant pour justifier un travail de recherche. Ce qui légitime le choix du sujet est le caractère omniprésent, incompris et pourtant crucial des outils d'inférence statistique dans la recherche biomédicale.

Il suffit de parcourir les résumés des articles analysant des données expérimentales pour se rendre compte de l'importante prévalence de ces outils dans la recherche. Ils sont omniprésents mais, paradoxalement, très mal compris par bien des scientifiques. Parmi les observateurs de la littérature (biomédicale notamment), nombreux sont ceux qui dénoncent la manière dont ces notions statistiques sont utilisées. Thompson, par exemple, recensait, en 2001, pas moins de 402 articles remettant en cause la manière dont les outils d'inférence statistique sont utilisés dans la littérature scientifique [Thompson, 2001] [1].

Malheureusement, l'incompréhension et les usages abusifs de ces outils ont des conséquences regrettables dans la construction des connaissances scientifiques.

En effet, ils interviennent dans l'étape – cruciale – de la confrontation des données ex-

1. A titre d'évocation, voici quelques titres de cette série d'articles : "*Null hypothesis testing : problems, prevalence, and an alternative*", "*Potential pitfalls in the use of p-values in the interpretation of significance levels*", "*Statistical analysis and the illusion of objectivity*", "*Uses and misuses of hypothesis testing*", "*The case against statistical significance testing*", "*Statistical significance tests : scientific ritualism or scientific method ?*" , "*Some problems connected with statistical inference*", "*In criticism of the null hypothesis statistical test*", "*Significance tests die hard : the amazing persistence of a probabilistic misconception*", "*What's the difference ? Pediatric residents and their inaccurate concepts regarding statistics*", *etc.*

périmentales avec les hypothèses préalablement engagées dans la recherche. En cela, les outils d'inférence statistique se trouvent au cœur de l'activité du scientifique. En outre l'augmentation rapide de la quantité de données disponibles pour les scientifiques ces dernières années risque encore d'amplifier le problème, l'écart se creusant entre les compétences nécessaires pour analyser des quantités de données toujours plus grandes et les compétences réelles des chercheurs.

> "*More people have more access to data than ever before. But a comparative lack of analytical skills has resulted in scientific findings that are neither replicable nor reproducible. It is time to invest in statistics education*" [Peng, 2015].

Il est aujourd'hui avéré que l'incompréhension et les mésusages des outils d'inférence statistique par les scientifiques ont contribué à la *crise de reproductibilité de résultats* qui traverse la communauté scientifique [Wasserstein and Lazar, 2016, Nuzzo, 2014, Baker, 2016].

Améliorer la manière dont les scientifiques sont formés aux outils d'inférence statistique semble donc être un enjeu majeur et passera, notamment, par une réflexion en profondeur sur l'enseignement.

Mais comment procéder ? Comment peut-on améliorer l'enseignement de l'inférence statistique ?

Telle est la question générale sous-tendant notre recherche.

Dans un premier temps, nous présenterons ce qui constitue notre point de départ, à savoir les travaux réalisés antérieurement dans notre unité de recherche. Nous aborderons brièvement les principales orientations et les principaux résultats de ces travaux afin d'exposer, ensuite, les éléments de continuité et les points de rupture entre la présente thèse et ces travaux antérieurs. Enfin, nous esquisserons le plan général de notre travail.

Mais, avant tout, commençons par définir l'expression "inférence statistique" qui est au cœur de ce travail et ne va pas de soi pour le lecteur non-statisticien.

1.1 Statistique et inférence statistique

De manière très générale, la **statistique** est la discipline qui s'intéresse à la collecte, à l'analyse et à l'interprétation des données [Dodge, 2003]. La **biostatistique** est l'application des méthodes de la statistique aux données biologiques et médicales (d'après [Armitage and Colton, 1998] cité dans [Dodge, 2003]).

Les buts poursuivis par le statisticien lorsqu'il collecte, analyse et interprète des données sont bien décrits par Romeijn (2017) :

> *"Statistics is a mathematical and conceptual discipline that focuses on the relation between data and hypotheses. The data are recordings of observations or events in a scientific study, e.g., a set of measurements of individuals from a population. The data actually obtained are variously called the sample, the sample data, or simply the data, and all possible samples from a study are collected in what is called a sample space. The hypotheses, in turn, are general statements about the target system of the scientific study, e.g., expressing some general fact about all individuals in the population. A statistical hypothesis is a general statement that can be expressed by a probability distribution over sample space, i.e., it determines a probability for each of the possible samples"* [Romeijn, 2017].

La statistique se divise généralement en deux champs principaux : l'analyse descriptive et l'inférence statistique.

L'analyse descriptive est définie comme :

— "*The use of statistics to describe a set of known data in a clear and concise manner, as in terms of its mean and variance, or diagrammatically, as by a histogram*" [Collins English Dictionary, 2020] ;

— "[Descriptive statistics] are *methods of analysis, graphical or tabular, without any probabilistic formulation.* " [Dodge, 2003].

L'inférence statistique, quant à elle, reçoit les définitions suivantes :

— *"Inferential statistics : The theory, methods, and practice of forming judgments about the parameters of a population, usually on the basis of random sampling"* [Collins English Dictionary, 2020] ;

— *"Inferential statistics : The process of making inferences about a population from findings based on sampled observations. Inferential statistics are used to go beyond the description of the data and to examine hypotheses about underlying research questions."* [Dodge, 2003] ;

— "[Statistical inference] *moves beyond the data at hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are uncertain "*([Moore, 2007, p.172], cité dans [Park, 2012]). ;

— *"A statement about statistical populations made from given observations with measured uncertainty. An inference in general is an uncertain conclusion. Two things mark out statistical inference. First, the information on which they are based is statstical, i.e., consists of observations subject to random fluctuations. Secondly, we explicitly recognise that our conclusion is uncertain, and attempt to measure, as objectively as possible, the uncertainty involved"* [Cox, 1958].

Sur base de ces définitions, et eu égard aux points que nous aborderons, nous définirons les

concepts comme suit :

L'**analyse descriptive** est un exercice de synthèse dont l'objectif est de rendre intelligible des observations en les résumant le plus possible tout en conservant un maximum de l'information initiale.

L'**inférence statistique** est un exercice de généralisation dans lequel on utilise les résultats observés dans un échantillon pour dire "quelque chose" à propos de la population hypothétique de laquelle cet échantillon provient. Pour y parvenir, on pose comme hypothèse que les échantillons de cette population se distribuent selon une certaine loi de probabilité définie par un ou plusieurs paramètre(s) inconnu(s).

Dans ce cadre, il existe, classiquement, deux approches concernant les paramètres inconnus.

D'une part, on peut chercher à déterminer, sur bases d'observations qui auraient été faites sur un échantillon, la valeur plausible ou les valeurs plausibles pour ces paramètres. Les outils statistiques qui permettent de répondre à ces questions sont classés dans les **outils d'estimation** parmi lesquels on retrouve l'estimation ponctuelle, l'intervalle de confiance ou encore l'intervalle de crédibilité.

D'autre part, on peut chercher à se prononcer à propos d'hypothèses statistiques qui concernent la valeur du paramètre étudié. Il existe, pour cela, différents **outils de test** qui répondent à des questions différentes, parmi lesquels on citera :

- le test de significativité (selon Fisher) qui cherche à mesurer le niveau d'accord entre les observations et une certaine hypothèse statistique, et cela à travers la P-valeur ;
- le test d'hypothèses (selon Neyman et Pearson) qui cherche à déterminer, sur base des observations, l'hypothèse statistique qu'il convient de considérer comme vraie parmi plusieurs hypothèses concurrentes ;
- le test de significativité (selon Jeffreys) qui cherche à mesure à quel point les observations soutiennent plus une hypothèse A qu'une hypothèse B à travers le facteur de Bayes.

1.2 Point de départ

Ce travail de thèse n'a pas été conçu *ex nihilo* : il s'appuie, entre autres, sur des travaux qui ont été menés sur le même thème au sein de notre unité de recherche [2]. Dès lors, pour introduire l'approche que nous développons dans cette thèse, nous jugeons utile de présenter brièvement ces travaux antérieurs afin de montrer les points de continuité et de rupture.

Confronté aux problèmes, locaux, de l'enseignement de l'inférence statistique, Philippe Calmant (2004) met au point un dispositif didactique novateur pour l'époque. En effet, il intègre le

2. Unité de Recherche en Didactique de la Biologie, URDB.

cours théorique à une plate-forme Web permettant aux étudiants d'avoir accès à des résumés de la matière et à des questionnaires d'exercices corrigés en temps réel et proposant une navigation modulaire entre les différents thèmes abordés. Le dispositif d'enseignement sera, par la suite, complété par une plateforme d'auto-évaluation formative, *eTests* [Vincke et al., 2014].

Globalement, la mise en place du site d'auto-apprentissage aura pour effet de favoriser l'autonomie des étudiants dans l'apprentissage des notions de statistique dont le test d'hypothèses. Il permettra également, selon Philippe Calmant, d'alléger l'encadrement des séances d'exercices. Cependant, il ne parviendra pas à résoudre le problème de l'enseignement de l'inférence statistique [Calmant, 2004].

Dans sa thèse, Philippe Calmant propose une analyse didactique des difficultés des étudiants envers le test d'hypothèses en se focalisant sur le public de deuxième année d'études universitaires en biologie. Cette analyse repose sur la théorie des situations didactiques de Brousseau et sur la théorie anthropologique du didactique de Chevallard. Malgré l'utilisation du nouveau dispositif didactique, deux difficultés majeures persistent pour les étudiants : (1) la prise en compte de la variabilité dans les raisonnements liés au test d'hypothèses et (2) la lecture de la courbe de Gauss.

Les perspectives de cette thèse appellent à un retour à un dispositif d'enseignement plus concret, ne reposant plus exclusivement sur le numérique. Les recherches qui suivent la thèse de Philippe Calmant consistent à mettre au point une séance de travaux pratiques dans laquelle les étudiants manipulent et mesurent des objets concrets, physiques (des galets de taille variable) et décrivent ceux-ci en construisant des graphiques à partir de blocs de construction [Vincke et al., 2013]. Cette séance s'oriente donc principalement autour de notions liées à l'analyse descriptive (variance et histogramme) et, s'il est possible qu'elle ait un impact sur la compréhension de ces notions-là, les difficultés des étudiants au moment d'aborder l'inférence statistique demeurent.

Le présent travail s'inscrit dans la continuité de ces travaux mais également dans une certaine rupture.

Notons, tout d'abord, que le **contexte professionnel** est resté sensiblement le même. Celui-ci comprend un professeur de biostatistique et une équipe d'environ six assistants impliqués dans l'enseignement des notions de statistique aux étudiants universitaires issus de différentes sections (Biologie, Géographie, Géologie, Sciences biomédicales, Pharmacie, Médecine et Médecine vétérinaire) à l'Université de Namur. Cette équipe enseignante fait partie de l'Unité de Recherche en Didactique de la Biologie (URDB), unité qui s'intègre dans le département de Biologie au sein de la faculté des Sciences.

A l'instar des travaux de Philippe Calmant, nous nous focaliserons sur **la question de l'enseignement de l'inférence statistique**. Cependant, là où il considère la matière enseignée,

le test d'hypothèses en l'occurrence, comme une donnée "non-négociable", nous tenterons de prendre du recul par rapport au savoir enseigné en le comparant, d'une part au savoir "savant" et, d'autre part, à ce que l'on peut appeler les *pratiques sociales de référence* [Martinand, 1989].

Concernant le **cadre théorique**, nous nous baserons également sur la théorie des situations didactiques [Brousseau, 1997] et sur la théorie anthropologique du didactique [Chevallard, 1991]. Nous utiliserons la première pour étudier les connaissances qui peuvent faire obstacle à l'apprentissage du test d'hypothèses chez des étudiants soumis à certaines situations et la seconde pour questionner le savoir enseigné et le replacer au sein d'un contexte institutionnel soumis à certaines contraintes.

On peut noter également de légères différences concernant la définition du **public cible**.

Philippe Calmant s'intéresse aux étudiants de deuxième candidature [3] en Biologie. Dans notre travail, nous nous focaliserons sur les étudiants bacheliers en Sciences biomédicales, Pharmacie et Médecine. On aurait pu inclure les étudiants en Biologie dans notre public cible (ainsi que ceux en Médecine vétérinaire) dans la mesure où le problème de l'incompréhension des outils d'inférence statistique existe aussi au niveau de la recherche en biologie (voir [Thompson, 2001]). Nous nous centrerons ici sur un public relativement homogène et dont la formation en statistique est assez brève (voir tableau 1.1), là où elle s'étire sur plusieurs unités d'enseignement chez les étudiants en Biologie.

TABLE 1.1 – **Aperçu de la formation en biostatistique pour les étudiants en bachalauréat dans une des filières biomédicales à l'université de Namur pour l'année académique 2019-2020**.

Section	**Cours**	**Crédits**	**Théorie**	**Exercices**
Biologie	Introduction aux notions statistiques (SBIOB205)	3 ECTS	15h	15h
	Statistiques avancées en sciences de la vie (SBIOB132)	4ECTS	30h	30h
Sciences biomédicales	Introduction aux notions de statistique médicale (MMEDB283)	4 ECTS	24h	15h
Pharmacie	Introduction aux notions de statistique médicale (MMEDB283)	4 ECTS	24h	15h
Médecine	Biostatistique (MBIOB131)	4 ECTS	24h	15h

ECTS : *European Credit Transfer and Accumulation System.*

Si le public est assez proche, on observe des différences plus fondamentales en ce qui concerne

3. Les deux années de candidatures correspondent, approximativement, aux trois années du bachelier actuel.

la portée du travail. Philippe Calmant cherche à améliorer l'enseignement à un niveau *local* ("Comment améliorer la compréhension des biostatistiques pour les étudiants de deuxième candidature en sciences en exploitant les outils d'enseignement issus du monde des multimédias ?"), là où nous chercherons plutôt à éclairer un problème *global* ("l'incompréhension et les usages abusifs des outils d'inférence statistique en recherche biomédicale") à partir, notamment, d'une analyse de pratiques locales.

Au niveau *local*, on trouve les étudiants qui constituent notre public cible (voir plus haut) et le dispositif d'enseignement de la statistique que nous avons mis en place. A l'instar des travaux antérieurs, nous baserons notre analyse sur ce que l'on pourrait appeler, en statistique, un échantillon de convenance. En travaillant sur notre propre enseignement, nous avons facilement accès aux différents éléments qui constituent l'environnement didactique des apprenants et nous pouvons facilement expérimenter de nouveaux dispositifs. L'inconvénient est que l'on peut manquer de recul et, potentiellement, mélanger la posture de l'enseignant avec celle du chercheur. Il s'agit là d'une limite de notre approche dont nous discuterons au moment de tirer les conclusions de notre travail.

Le niveau *global* est celui des chercheurs, des professionnels déjà formés dans les disciplines liées au domaine biomédical. En effet, c'est bien parce qu'il existe un véritable problème lié à l'inférence statistique chez les chercheurs et que ce problème a des conséquences sur la manière dont le savoir scientifique se construit et sur la confiance que l'on peut lui accorder qu'il nous semble important de nous y intéresser. Le niveau professionnel, celui de la recherche biomédicale sera donc considéré comme un point de mire pour notre analyse : c'est lui qui justifie l'importance du sujet et c'est à ce niveau qu'il faudra juger les apports éventuels de notre travail. L'objectif ultime sous-jacent à notre recherche serait donc d'améliorer la recherche scientifique dans le domaine biomédical.

Par rapport aux travaux antérieurs, nous proposons de mettre de côté la dimension informatique, technologique et de donner une place plus importante à la recherche scientifique. Notre travail fera donc intervenir trois champs distincts : celui de la science, celui de la statistique et celui de la didactique, chacun jouant un rôle différent.

La science, et plus précisément la recherche scientifique dans le domaine biomédical, fournira le *but*. L'enseignement de l'inférence statistique aux futurs scientifiques du domaine biomédical a pour principal objectif de former des scientifiques capables de comprendre la littérature scientifique et surtout, capables de participer à la construction et à la diffusion de ce savoir. Il nous semble que c'est à l'aune de cet objectif que doivent être évalués les enseignements, y compris ceux liés à l'inférence statistique.

La statistique fournira *la matière*, les outils d'inférence statistique. Ils sont considérés ici comme des objets utilisés dans le but de construire du savoir scientifique. Pour bien comprendre en quoi ces outils permettent, ou ne permettent pas, de sereinement construire et diffuser des

connaissances, il nous faudra les analyser.

La didactique, quant à elle, fournira les *grilles d'analyses* et définira le champ d'étude. Dans ce travail nous nous focaliserons sur les questions liées à l'enseignement de l'inférence statistique et poserons un regard didactique sur les difficultés rencontrées par les futurs scientifiques dans la maitrise des outils d'inférence statistique.

1.3 Plan de la thèse

Dans cette thèse, nous partons donc d'une question générale : "Comment améliorer l'enseignement de l'inférence statistique aux futurs scientifiques du domaine biomédical ?". De celle-ci découlent deux questions plus spécifiques que nous présenterons successivement mais que nous avons développées en parallèle (voir figure 1.1).

D'une part, dans le chapitre 2, nous interrogerons le savoir enseigné au niveau local. Qu'enseigne-t-on et en quoi les notions que nous enseignons diffèrent-elles des notions décrites dans le savoir "savant" ?

Pour apporter des éléments de réponse à ces questions, nous nous appuierons sur le cadre théorique proposé par Chevallard dans sa théorie anthropologique du didactique. Ce cadre nous amènera étudier la *transposition didactique* de la démarche du test d'hypothèses. Au cours de cette analyse, nous tenterons de décrire le savoir "savant", le savoir enseigné ainsi que les pratiques sociales de référence en termes de *praxéologies*. La comparaison des praxéologies entre différentes institutions conduira à interroger les contraintes qui existent au sein de ces différentes institutions et qui définissent les *praxéologies* que l'on peut y trouver.

D'autre part, dans le chapitre 3, nous chercherons à décrire la nature des difficultés des étudiants à l'apprentissage du test d'hypothèses.

Pour ce faire, nous nous baserons sur le cadre conceptuel de la théorie des situations didactiques proposé par Brousseau ainsi que sur la méthodologie de l'ingénierie didactique défendue par Artigue [Artigue, 1989]. Nous y présenterons un dispositif expérimental consacré à l'enseignement du test d'hypothèses que nous utilisons sur des groupes d'étudiants universitaires des filières biomédicales. Nous tenterons de décrire les raisonnements des étudiants confrontés à ce dispositif expérimental et d'analyser leurs difficultés en termes d'obstacles.

En conclusion, dans le chapitre 4, nous ferons une synthèse des éléments qui découlent de ces analyses et nous verrons qu'une partie de ceux-ci sont ont un caractère qui dépasse le cadre local de l'analyse initiale. Nous discuterons des limites de notre approche et des pistes de solution que cette thèse nous invite à considérer.

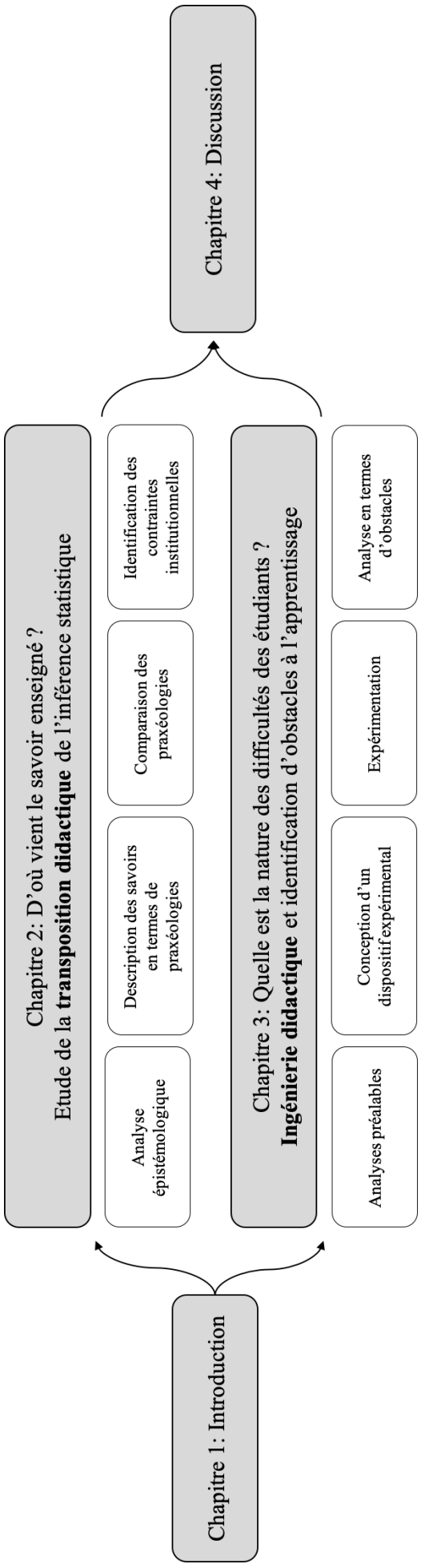


FIGURE 1.1 – **Représentation schématique de la structure générale de la thèse.**

Chapitre 2

Transposition didactique

2.1 Cadre théorique

Pour tenter de décrire la nature du savoir enseigné et voir en quoi celle-ci peut expliquer des difficultés d'apprentissages auxquelles nous sommes confrontés, nous allons, dans ce chapitre, nous appuyer sur la théorie anthropologique du didactique (TAD) de Chevallard.

Ce cadre théorique nous permettra de prendre du recul par rapport au savoir enseigné pour mieux pouvoir le remettre en questions.

> "*La TAD situe l'activité mathématique, et donc l'activité d'étude en mathématiques, dans l'ensemble des activités humaines et des institutions sociales. Or ce parti pris épistémologique conduit qui s'y assujettit à traverser en tous sens - ou même à ignorer - nombre de frontières institutionnelles à l'intérieur desquelles il est pourtant d'usage de se tenir, parce que, ordinairement, on respecte le découpage du monde social que les institutions établies, et la culture courante qui en diffuse les messages à satiété, nous présentent comme allant de soi, quasi naturel, et en fin de compte obligé.*" [Chevallard, 1998]

Le terme *institution* dans la théorie de Chevallard est à prendre dans un sens très large :

> "U*ne institution est un dispositif social "total" qui peut certes n'avoir qu'une extension très réduite dans l'espace social (il existe des "micro-institutions"), mais qui permet – et impose – à ses sujets, c'est-à-dire aux personnes qui viennent y occuper les différentes positions offertes dans l'institution, la mise en jeu de manières de faire et de penser propres" (Chevallard, 2003, p. 82). Selon ce point de vue, une classe, l'école, l'université, la famille ou encore la vie quotidienne, sont des institutions dans lesquelles différentes positions peuvent être occupées* [Koliopoulos et al., 2013].

Dans cette section nous allons donc présenter les éléments de la TAD qui nous semblent les plus pertinents pour notre recherche.

Nous partirons de la définition que Chevallard donne de la relation didactique. Nous verrons ensuite qu'il nous amène, dans un premier temps, à considérer, non pas *le* savoir mais *les savoirs* qui peuvent différer d'une institution à l'autre. Nous exposerons le concept de *praxéologie* qu'il propose afin de définir plus précisément ce qu'il entend par *savoir*. La décomposition d'un savoir en niveaux praxéologiques nous servira de grille d'analyse pour la comparaison du savoir en différentes institutions. Nous donnerons quelques exemples concrets d'analyses en termes de praxéologie afin de donner une idée de la manière dont l'auteur propose d'utiliser ce concept dans l'analyse didactique. Enfin nous verrons comment les concepts de praxéologie et de transposition didactique permettent à Chevallard de redéfinir l'objet de la didactique.

Une fois ces éléments présentés, nous serons en mesure d'énoncer plus précisément la question de recherche.

2.1.1 La relation didactique

Intuitivement, on pourrait penser que la relation didactique est une relation binaire entre un individu qui possède un savoir, le professeur, et un individu qui ne le possède pas, l'élève. Pour Chevallard, ce serait oublier un acteur majeur de la relation didactique : le savoir. Il décrit donc celle-ci comme une relation ternaire entre celui qui possède le savoir, celui qui ne le possède pas et le savoir en lui-même.

Selon lui, toutes les relations ternaires de ce type ne sont pas des relations didactiques. Par exemple, la relation entre le client et son garagiste, ou celle entre le patient et le médecin satisfont à la définition sans être des relations didactiques. Pour qu'il y ait relation didactique, il faut qu'il y ait une intention d'enseigner [Chevallard, 1989].

Cette intention d'enseigner ne doit pas être forcément considérée à l'échelle individuelle mais plutôt à l'échelle de la société. C'est la société, dans son ensemble, qui charge les acteurs de l'enseignement, regroupés au sein d'institutions, d'enseigner des savoirs aux élèves, à travers un contrat tacite.

Dans cette relation didactique, la spécificité de l'approche didactique, par opposition à l'approche pédagogique par exemple, est de s'intéresser de près au contenu enseigné, de le prendre pour objet d'étude. Sans réelle prise en compte du savoir dans cette relation ternaire, on obtient ce qu'il nomme une *réduction pédagogique*.

> *"[La réduction pédagogique] se fonde sur un postulat que le savoir étudié n'est pas problématique, que le problème tient tout entier dans le rapport des élèves au savoir, lequel à son tour est subordonné au rapport de l'élève au maître."* [Chevallard, 1997]

Pour le didacticien, le savoir est au cœur de la relation didactique, on ne peut comprendre la relation entre l'enseignant et l'apprenant qu'en s'intéressant de près au savoir.

> *"One might as well try to explain the relationship between the pianist and his audience, or the waiter and the customer, by ignoring the music or the food ! Certainly some facts can be explained on such a narrow basis, e.g. in terms of group dynamics."* [Chevallard, 1989]

> *"La didactique se donne ainsi un objet [le savoir] qui, dans la culture courante est ordinairement scotomisé entre deux états qui seuls comptent - l'état de non-savoir, l'état de savoir - les états didactiques sont un simple passage, nécessaire mais éphémère."* [Chevallard, 1997]

2.1.2 Du savoir aux savoirs

Pour Chevallard, le savoir est *problématique*, c'est-à-dire qu'il ne va pas de soi, contrairement à ce qu'on pourrait penser. En outre, le savoir ne se résume pas à deux états : *état de non-savoir* et *état de savoir*. L'auteur fait remarquer que, pour beaucoup de disciplines, notamment les mathématiques, le savoir enseigné diffère du savoir dit "savant" et que cela s'explique par le fait que savoir savant et savoir enseigné ont été construits ou reconstruits dans des *institutions* différentes qui répondent à des contraintes différentes.

> *"La question décisive, à cet égard, est celle de la problématicité des savoirs. Pour le pédagogue, les savoirs sont choses sûres. Or tout savoir n'est d'abord qu'une hypostase, une entité supposée, une idée de substance, dont nous faisons l'hypothèse en certains contextes institutionnels, en supposant que tel ou tel agit comme si ses gestes procédaient d'un certain corps de connaissances, que nous croyons deviner à travers son faire."* [Chevallard, 1997]

Or, dans nos sociétés, remarque-t-il, on rencontre souvent la conception selon laquelle il existerait *le* savoir et que ce savoir enseigné serait strictement identique au savoir savant, au savoir utilisé, ce qui participerait, d'ailleurs, à une certaine légitimation de cet enseignement. En effet, pourquoi enseignerait-on autre chose que le savoir savant ?

L'approche du didacticien consiste justement à questionner ce qui semble évident - l'unicité du savoir - ce qui lui permet d'étudier les écarts entre le savoir savant et le savoir enseigné. Le savoir savant est fait pour être utilisé, c'est-à-dire pour répondre à des questions.

> *"Bodies of knowledge are, with a few exceptions, not designed to be taught, but to be used. To teach a body of knowledge is thus a highly artificial entreprise. The transition from knowledge regarded as a tool to be put to use, to knowledge as something*

to be taught and learnt, is precisely what I have termed the didactic transposition of knowledge." [Chevallard, 1989]

"The first step in establishing some body of knowledge as teachable knowledge therefore consists in making it into a body of knowledge, i.e., into an organized and more or less integrated whole." [Chevallard, 1989]

"More generally, taught bodies of knowledge have been derived from corresponding scholarly bodies of knowledge, as I call them. Scholarly bodies of knowledge, in effect, tend to achieve a comparatively high degree of integration, in so far as they boast a mode of organization that I referred to earlier as theory – a mode of organization for which mathematics expressly provided the historically fundamental paradigm as expounded in Euclid's Elements." [Chevallard, 1989]

Le savoir tend donc à exister sous la forme d'un savoir savant. Mais, pour pouvoir être enseigné, pour exister au sein de l'institution scolaire, il doit d'abord être reconstruit au sein d'une structure compatible avec cette institution. En l'occurrence, il doit se présenter comme un ensemble cohérent dans lequel une progression linéaire est possible et qui est compatible avec le rythme et l'organisation scolaires.

Cette transformation du savoir savant en un objet d'enseignement est ce que Chevallard nomme la **transposition didactique**.

Une des activités du didacticien consiste en l'étude de cette transposition didactique, vue non pas comme un phénomène automatique, nécessairement réussi, mais plutôt à remettre en question inlassablement.

Le didacticien, en s'émancipant du rapport institutionnel au savoir, se donne les moyens de voir les contraintes spécifiques d'une institution, ce qui peut l'amener à imaginer d'autres savoirs enseignés, d'autres organisations pédagogiques que celle qui prévaut.

Le didacticien distingue, classiquement, quatre types de savoirs :

1. le *savoir savant*, que l'on retrouve dans le "monde académique" ;
2. le *savoir à enseigner* défini par une institution particulière composée d'enseignants de décideurs, de personnes influentes, ensemble que Chevallard nomme la *noosphère* ;
3. le *savoir enseigné*, que l'on retrouve dans le texte des supports de cours, dans le discours de l'enseignant, *etc.* ;
4. le *savoir appris* au niveau individuel.

"*In the case where no adequate scholarly body of knowledge exists, the intention to teach has often resulted in, or accompanied, an attempt to create a scholarly or,*

rather, a pseudo-scholarly body of knowledge, from which the intended taught knowledge could be shown to derive. (Accounting and its corresponding body of knowledge, accountancy, are a case in point.) These facts of counter-transposition speak well for the stability of the solution thus secured. The question remains however of why such a solution was, and still is, so widely embraced. Again, the explanation lies in the difference between used knowledge and taught knowledge. As long as you only use knowledge in doing something, you need not justify nor even acknowledge the used knowledge in order to endow your activity with social meaning." [Chevallard, 1989]

Chevallard imagine également l'existence possible d'une **contre-transposition didactique** qui prendrait place lorsque le savoir savant correspondant au savoir enseigné n'existerait pas en tant que tel. Par extension, le concept de transposition didactique sera élargi à toutes les institutions, qu'elles soient liées au système d'enseignement ou non, et Chevallard parlera, de manière plus générale, de **transposition institutionnelle**.

Ce concept de transposition institutionnelle permet notamment d'interroger les écarts que l'on peut trouver entre le savoir savant et ce que Martinand nomme les **pratiques sociales de référence**, c'est-à-dire des pratiques propres à un secteur social, qui diffèrent des pratiques scolaires et qui lui servent de référence. C'est notamment le cas de la recherche scientifique, des pratiques industrielles, d'activités culturelles, *etc.* [Martinand, 1989]

2.1.3 Des savoirs aux praxéologies

Avec le concept de transposition didactique, Chevallard nous invite à considérer qu'il existe non pas un savoir mais des savoirs, qui sont fonction des contraintes des institutions dans lesquelles ils existent.

Dans sa théorie anthropologique du didactique, Chevallard poursuit sa dissection du concept de savoir avec l'idée qu'il n'existe pas des savoirs, objets qui existeraient indépendamment des institutions et des individus, mais des *rapports au savoir*, personnels ou institutionnels.

"Il y a par exemple la notion supposée de logarithme, qu'aucune personne ni aucune institution ne saurait "posséder" ; et il y a le rapport que j'ai, personnellement à cette notion, comme il y a le rapport que l'on devrait avoir à elle quand on occupe légitimement telle position en telle institution - rapport qui, au lycée, ne sera pas le même pour le professeur de mathématiques et pour le professeur de physique et chimie par exemple. " [Chevallard, 2007].

Il est donc possible de déceler et d'étudier des manières de savoir chez des individus ou des institutions, de détailler des rapports aux savoirs.

Chevallard propose le concept d'*organisation praxéologique* ou, simplement, de **praxéologie** pour décomposer les différentes manières de savoir, les relations qu'un sujet ou une institution peuvent avoir vis-à-vis d'un savoir. L'étymologie du mot *praxéologie* fait référence aux deux composantes que l'on considère généralement derrière le terme savoir : le savoir-faire (praxis) et le savoir théorique (logos). Le premier correspondrait plutôt, selon Chevallard, à l'association tâche/technique tandis que le second correspondrait à la paire technologie/théorie.

> " *Toujours le savoir fait problème. Au-delà d'un faire observable, un savoir supposé renvoie, plus globalement, à ce que je nomme une* organisation praxéologique*, ou* praxéologie*. A l'origine d'une praxéologie se trouvent une ou plusieurs* questions*qui, génétiquement, apparaissent comme les* raisons d'être *de l'organisation praxéologique, parce que celle-ci est censée leur* apporter réponse*. Dans l'immense majorité des cas, on peut ramener ces questions à des formulations du type "Comment faire pour... ?". En d'autres termes, une praxéologie doit nous permettre d'accomplir certaines* tâches*, les tâches* t *d'un certain* type *de tâches* T." [Chevallard, 1997]

Une praxéologie est définie par quatre composantes : tâche, technique, technologie et théorie.

Au départ, une praxéologie se construit en réponse à une question, souvent du type "Comment faire pour... ?". La **tâche** est donc ce que la praxéologie nous permet d'accomplir. Une manière de faire, de réaliser la tâche, est appelée une **technique**. C'est une réponse à la question "comment fait-on pour...". Cette manière de faire est justifiée par un discours qui l'éclaire, la **technologie**.Enfin, ce discours est, lui-même, soutenu par une **théorie** plus générale.

La décomposition d'une réponse à une certaine tâche en trois niveaux (technique/technologique et théorique) permet, selon Chevallard, de décrire adéquatement les savoirs.

> *"Bien entendu, on peut imaginer que cette régression justificative se poursuive à l'infini - qu'il y ait une théorie de la théorie, etc. En fait, la description à trois niveaux présentée ici (technique/technologique/théorie) suffit, en général, à rendre compte de l'activité à analyser. La théorie, terre d'élection des truismes, tautologies et autres évidences, est même souvent évanouissante : la justification d'une technologie donnée est, en bien des institutions, traitée par simple renvoi à une autre institution, réelle ou supposée, censée détenir une telle justification. C'est là le sens du classique "On démontre en mathématiques..." du professeur de physique, ou encore du "On a vu en géométrie..." du professeur de mathématiques d'autrefois."* [Chevallard, 1998]

2.1.4 Exemples d'analyse praxéologique

Voici deux exemples illustrant la manière dont Chevallard envisage une analyse praxéologique.

Comparaison de fractions

L'auteur donne l'exemple suivant. Imaginons que l'on cherche comment déterminer si deux fractions sont égales : par exemple 9/14 et 279/434. Une manière de répondre à la question serait d'entrer 9/14 et 279/434 dans la calculatrice et d'observer que $9/14 = 0.6428\ldots$ et que $279/434 = 0.6428\ldots$ et d'en conclure à l'égalité des fractions.

Or cette technique pourrait être rejetée sur base du principe théorique général selon lequel "*les suites de décimales pourraient différer plus loin, au-delà de la 12^e décimale, voire au-delà de la 30^e décimale par exemple*"[Chevallard, 2007]. Pour pouvoir l'appliquer, il faut la justifier par une technologie. Dans le cas présent, cette technologie pourrait prendre la forme suivante :

> *"En fait, comme 434 est un multiple de 14, il y a égalité dès que la différence est en valeur absolue inférieure à* $1/434 \leq 0,0023$*, en sorte qu'il suffit de savoir que les fractions* $a/14$ *et* $b/434$ *ont leurs trois premières décimales identiques pour pouvoir conclure à leur égalité. Le précepte "théorique" qui, en ce cas, engendre couramment le rejet d'une technique pourtant parfaitement justifiable est bien connu : "se méfier de la calculatrice." En mathématiques comme en tout domaine d'activité, c'est ainsi au niveau théorique que se révèlent à qui sait les lire les limitations et les failles de la connaissance d'une institution."* [Chevallard, 2007]

Dans l'exemple, la tâche consiste à conclure à l'égalité ou non de deux fractions a/b et c/d. La technique consiste à les encoder dans la calculatrice et à comparer leurs trois premières décimales. La technologie démontre que, en effet, la différence entre ces deux fractions ne pourrait être plus grande que $1/(b \times d)$ si a,b,c et d appartiennent à N*. La théorie n'apparait pas de manière explicite dans l'exemple mais elle est censée contenir tous les principes utilisés pour justifier la technologie. De plus, l'exemple montre les principes théoriques utilisés pour rejeter, dans un premier temps, la technique proposée sans justification : "Les suites de décimales pourraient toujours différer plus loin" et "Il faut se méfier de la calculatrice".

Comparer des nombres décimaux

Un autre exemple est fourni par Chevallard (2013) et concerne l'analyse d'extraits d'un livre de référence décrivant le savoir à enseigner, relatif à la comparaison de nombres décimaux.

— "*De deux nombres décimaux, le plus grand est celui qui a la plus grande partie entière.*

— *Si les parties entières sont égales, on compare les parties décimales, décimale par décimale. On compare les chiffres des dixièmes, puis, s'ils sont égaux, les chiffres des centièmes, etc. Comparaison de* $8,1\boldsymbol{6}9$ *et* $8,1\boldsymbol{4}023$ *:* $6 > 4$ *donc* $8,169 > 8,14023$.

— *On peut aussi comparer les parties décimales globalement. On commence alors par réécrire les nombres avec le même nombre de décimales. Comparaison de* $2,01$ *et* $2,013$ *:* $2,01 = 2,010$. *Comparer* $2,\boldsymbol{010}$ *et* $2,\boldsymbol{013}$ *revient à comparer* 10 *et* 13. $10 > 13$ *donc* $2,01 < 2,013$ *(p.58).*"

(Le collège en poche. Tout le programme de 6^e en fiches, (Maxi-Livres, 2002), cité dans [Chevallard, 2013]).

De là, il identifie le type de tâche T : "*comparer deux nombres décimaux (différents), c'est-à-dire déterminer quel est le plus grand et quel est le plus petit.*"

Il analyse ensuite la praxéologie relative à ce type de tâche. Celle-ci semble contenir plusieurs éléments techniques et technologiques :

Le premier "*principe technologique* qui guide et justifie un certain *geste technique* (à portée limitée)" qu'il relève est l'idée selon laquelle si deux nombres décimaux ont des parties entières différentes, alors le plus grand est celui qui a la plus grande partie entière. Si les parties entières sont identiques, alors il convient d'examiner les parties décimales. Pour cela, on retrouve deux techniques différentes : τ_1 et τ_2.

La technique τ_1 consiste à procéder décimale par décimale, en partant du premier au dernier chiffre après la virgule. La justification technologique de cette technique n'apparait pas dans le texte étudié, Chevallard propose que la technologie θ_1 suivante pourrait justifier la technique τ_1.

1. *Les écritures décimales a=3,46... et b= 3,41... (par exemple) désignent respectivement les nombres sommes :* $a = 3 + \frac{4}{10} + \frac{6}{100} + \ldots$ *et* $b = 3 + \frac{4}{10} + \frac{1}{100} + \ldots$;
2. *Le nombre somme* $a* = 3 + \frac{4}{10} + \frac{6}{100}$ *est (à l'évidence) strictement supérieur au nombre somme* $b* = 3 + \frac{4}{10} + \frac{1}{100}$;
3. *Les sommes "restes" (notées ci-dessus* $+\ldots$ *dans l'écriture de a et de b) sont toujours strictement inférieures à* $\frac{1}{100}$ *et ne peuvent donc pas modifier le résultat obtenu par la comparaison de* $a*$ *et de* $b*$ *: a et b sont rangés dans le même ordre que* $a*$ *et* $b*$.

[Chevallard, 2013]

La technique τ_2 consiste à procéder globalement, en comparant des suites décimales de même

longueur. Par exemple, pour comparer les chiffres 7,2348 et 7,235, on réécrira ce dernier sous la forme 7,2350 et puis on comparera les parties décimales entre elles : $2348 < 2350$.

Pour Chevallard, cette manière de procéder ne laisse pas transparaître de justification technologique et donc s'approche d'une *recette* ou d'un *dogme*. S'il fallait en trouver une, la justification technologique θ_2 pourrait avoir, selon lui, la forme suivante :

1. *Les écritures décimales* $a = 3,407$ *et* $b = 3,41$ *(par exemple) désignent respectivement les fractions successives suivantes :*
   $a = \frac{3407}{1000} = \frac{34070}{10000} = \frac{34700}{100000} = \ldots$ et $b = \frac{341}{100} = \frac{3410}{1000} = \frac{34100}{10000} = \ldots$;
2. *De deux fractions ayant le même dénominateur, la plus grande est celle qui a le plus grand numérateur ; il en résulte que, en l'espèce, on a :*
   $a = 3,407 = \frac{3407}{1000} < \frac{3410}{1000} = 3,14 = b$.

[Chevallard, 2013]

Ces techniques ont pour objet l'évitement d'une erreur courante qui consiste à comparer des parties décimales en les regardant comme des entiers. Par exemple $3,41 < 3,407$ parce que $41 < 407$. Et cette erreur semble être causée par une *difficulté d'origine théorique* liée à la manière dont nous lisons actuellement les nombres décimaux : 3, 41 se lit "trois virgule quarante et un" et 3, 407 se lit "trois virgule quatre cent et sept".

Or, il n'en n'a pas toujours été ainsi. Chevallard compare la praxéologie proposée dans ce livre de référence à une autre praxéologie proposé dans un manuel plus ancien dans lequel on peut lire :

> ***Pour lire un nombre décimal écrit.*** *- Pour lire un nombre décimal, on énonce d'abord la partie entière qu'on fait suivre du mot* ***entiers*** *ou* ***unités****, puis la partie décimale, comme s'il s'agissait d'un nombre entier, en la faisant suivre du nom des unités que représente le dernier chiffre décimal.*
>
> *Par exemple :* $4,075$ *se lit :* quatre unités soixante-quinze millièmes ; $25,00317$ *se lit :* vingt-cinq unités trois cent dix-sept cent-millièmes. (Bourlet (1922), cité dans [Chevallard, 2013]).

Ainsi, Chevallard compare les deux praxéologies et note qu'elles semblent sous-tendues par des principes théoriques différents. Derrière l'ancienne praxéologie on trouverait :

> "*une théorie "réaliste" des nombres : on n'y parle pas du nombre un mais de l'unité ; et, dans cette perspective, on parle de dixième, de centième, etc., de l'unité. Il semble qu'il n'y ait rien de tel dans le cas de la fiche examinée : les nombres semblent ne renvoyer qu'à leur écriture formelle.*" [Chevallard, 2013].

Dans la pratique plus récente, les nombres semblent avoir perdu ce lien avec une mesure de grandeur et ne renvoient à rien d'autre qu'à eux-mêmes. La manière dont les nombres décimaux sont lus a, elle aussi, évolué, s'est accélérée en perdant le lien avec des mesures concrètes. Cela a fait naître la possibilité d'une erreur dans la comparaison de nombres décimaux, celle qui consiste à prendre les parties décimales pour des entiers et à les comparer mêmes si elles sont de longueurs différentes. Les techniques τ_1 et τ_2 sont donc là pour éviter cette erreur.

2.1.5 Une définition de la didactique

A travers sa théorie anthropologique du didactique, Chevallard révèle sa vision de ce que devrait être la didactique. Il réagit notamment à la conception selon laquelle les écoles seraient un lieu où *le* savoir se *transmet*.

> *"L'usage s'est imposé, en français, de parler de la transmission d'un savoir, comme si un savoir était une réalité ou entièrement matérielle (tel un bien que l'on cède à autrui), ou entièrement symbolique (comme l'est un droit dont on hérite). (...) une organisation praxéologique ne saurait être simplement "transmise". Même lorsque cette organisation existe déjà en mille institutions, on ne saurait la "transporter" en une nouvelle institution à la manière dont on déménage un meuble - par simple transfert. Il convient au contraire de l'y reconstruire, de la recréer en cet habitat nouveau, à l'écologie peut-être fort différente."* [Chevallard, 1997]

> *"De là qu'on parle, en didactique, depuis bientôt vingt ans, de transposer un savoir, au sens quasi musical du terme - "faire passer (une forme musicale) dans un autre ton sans l'altérer"-, et non de le "transférer" ou de le "transmettre". Le mot de transposition désigne ainsi non une pratique toute constituée, et garantie, mais un grand problème, indéfiniment ouvert : comment "faire passer"dans un autre "ton institutionnel" sans "altérer" ? Ou du moins sans trop altérer, en contrôlant les altérations nécessairement imprimées."* [Chevallard, 1997]

La didactique ne serait donc pas l'étude de la manière dont le savoir se transmet en milieu scolaire mais plutôt l'étude de l'évolution des praxéologies au sein des institutions humaines. Par le recul vis-à-vis de son objet d'étude, la didactique s'autorise la remise en question des praxéologies institutionnelles et les décrit non pas comme des points de référence intangibles mais comme des points d'équilibre dans un système soumis à certaines contraintes qu'il convient de mettre en évidence si on veut pouvoir s'en défaire et imaginer d'autres modes d'organisation.

> *"La didactique se voue à étudier les conditions et contraintes sous lesquelles les praxéologies se mettent à vivre, à migrer, à changer, à opérer, à dépérir, à disparaître, à renaître, etc., au sein des institutions humaines."* [Chevallard, 2007]

2.2 Question de recherche

L'observation de difficultés récurrentes pour les étudiants lors de l'enseignement de l'inférence statistique nous amène à interroger le savoir enseigné : d'où vient-il ? Quels liens entretient-il avec le savoir savant ainsi qu'avec les pratiques sociales de référence ?

Pour apporter des éléments de réponse à ces questions, nous nous appuyons sur la théorie anthropologique du didactique de Chevallard. Celui-ci suggère d'analyser de manière critique le savoir enseigné et propose de le regarder, non pas comme une entité qui existerait indépendamment des institutions humaines et qui serait, nécessairement, équivalente au sein d'institutions différentes, mais plutôt comme un objet qui s'adapte aux contraintes institutionnelles, contraintes qui peuvent modifier considérablement la nature du savoir.

Le savoir que nous chercherons à décrire sera constitué des outils d'inférence statistiques fonctionnant sous la forme de test statistique.

Pour décrire la nature du savoir enseigné et voir en quoi il diffèrerait des versions du savoir que l'on peut trouver en d'autres institutions, nous utiliserons la notion de *praxéologie* décrite par Chevallard. Cette manière de disséquer le savoir en différentes entités (tâche, technique, technologie et théorie) nous servira de grille d'analyse pour comparer les différentes versions du savoir existant en différentes institutions.

Plus précisément nous allons tenter de décrire trois types de savoirs :
- le savoir savant que l'on trouvera dans les publications des auteurs ayant développé les théories relatives aux tests statistiques ;
- le savoir enseigné que l'on décrira principalement à partir du contenu des différents supports utilisés dans le cadre de notre enseignement ;
- les pratiques de référence qui sont constituées, dans notre cas, des pratiques d'utilisation des tests statistiques dans le domaine de la recherche biomédicale.

Selon Chevallard, il est normal que les praxéologies qui existent dans le savoir savant et dans le savoir enseigné diffèrent car elles ne sont pas soumises aux mêmes contraintes.

La question de recherche de ce premier chapitre est donc la suivante :

Quelles contraintes déterminent les diverses praxéologies existant au niveau du savoir enseigné ?

Pour tenter de répondre à cette question, nous suivrons le plan suivant :

Dans un premier temps, nous nous intéresserons à trois praxéologies relatives aux tests statistiques que l'on peut rencontrer au sein du savoir savant et qui nous semblent les plus proches du savoir enseigné.

— praxéologie A : le test de significativité selon Fisher (et décrit dans son livre *Statistical Methods for Research Workers* de 1925) ;
— praxéologie B : le test d'hypothèses selon Neyman et Pearson (décrit dans les articles *On the problem of the most efficient tests of statistical hypotheses* de 1933 et *Outline of a theory of statistical estimation based on the classical theory of probability* de 1937) ;
— praxéologie C : un test statistique bayésien utilisant le concept de probabilité *a posteriori* (basé sur des sources secondaires : Altman (1994) et Berry (2006)).

Dans un deuxième temps, nous nous intéresserons à la praxéologie enseignée à l'Université de Namur aux étudiants des filières biomédicales (praxéologie D). Nous verrons que cette praxéologie D semble mélanger des caractéristiques de A et de B. Nous tenterons d'identifier les contraintes institutionnelles propres à D et nous verrons que ces contraintes ne peuvent expliquer à elles seules la transposition didactique que nous avons décrite.

Pour la comprendre, il faudra s'intéresser à ce qui se passe au niveau des pratiques sociales de référence.

Dans un troisième temps, nous décrirons donc les pratiques sociales de référence, c'est-à-dire les pratiques liées au test statistique que l'on rencontre au sein de la littérature de recherche biomédicale. Nous y trouverons des concepts statistiques qui ressemblent aux concepts décrits dans le savoir savant mais qui y ont été transformés. Le thème statistique est donc le même mais les praxéologies diffèrent. A partir des écarts couramment dénoncés entre le savoir savant et les pratiques de référence, nous tenterons de définir une cinquième praxéologie (E) qui correspond à ces pratiques dénoncées. La comparaison entre les cinq praxéologies ainsi définies mettra en évidence que celle enseignée (D) s'approche bien plus de celle pratiquée (E) que de celles que l'on rencontre dans le savoir savant (A et B). L'analyse en termes de contraintes institutionnelles nous permettra de mieux expliquer l'importante transposition institutionnelle qui a eu lieu entre le savoir savant et les pratiques de références et donc, indirectement, entre le savoir savant et le savoir enseigné.

2.3 Savoir savant

2.3.1 Test de significativité (A)

Dans son livre *Statistical Methods for Research Workers* [Fisher, 1925], Ronald Fisher formalise des pratiques qui avaient déjà été énoncées dès 1900 par Karl Pearson (le test du χ^2) et dès 1904 par William Sealy Gosset [1] (la distribution du t de Student). Cependant son livre aura un retentissement bien plus important ce qui a pour conséquence que l'on associe généralement

1. Alias *Student*

son nom au test de significativité.

Il fournit, aux chercheurs de l'époque, un état des lieux de la manière dont la théorie statistique peut être utilisée dans le domaine de la biologie. Chaque concept statistique est donc présenté et accompagné de nombreux exemples concrets.

Pour tenter de décrire la praxéologie A, liée au test de significativité nous allons nous intéresser de plus près à l'exemple qu'il donne pour expliquer l'utilisation de la distribution du χ^2 pour comparer des effectifs observés à des effectifs attendus sous un modèle.

Il s'agit d'un exemple de génétique impliquant deux traits de caractère de la plante *Primula*[2] à savoir :

— le type de feuille : plate (allèle dominant) ou ondulée (allèle récessif)

— l'aspect du centre de la fleur : forme un "normal eye" (allèle dominant) ou un "Primrose Queen eye" (allèle récessif).

La question posée concerne la répartition des allèles dans des plantes obtenues par croisement de la génération hybride (F1). Si la répartition de ces deux traits est indépendante, on s'attend à une répartition mendeléenne (9 :3 :3 :1) de ces allèles. La question est donc de savoir si les données empiriques sont en accord avec l'hypothèse de la répartition mendeléenne de ces traits (voir tableau 2.1).

TABLE 2.1 – **Valeurs observées et attendues sous l'hypothèse d'une répartition mendeléenne de deux traits de caractère chez la primevère** .

	Flat Leaves.		Crimped Leaves.		Total.
	Normal Eye.	Primrose Queen Eye	Lee's Eye.	Primrose Queen Eye.	
Observed $(m+x)$	328	122	77	33	560
Expected (m) .	315	105	105	35	560
x^2/m . . .	·537	2·752	7·467	·114	10·870

Source : [Fisher, 1925]

Cependant on sait que, même si la répartition de ces deux traits était réellement[3] de 9 :3 :3 :1, les proportions observées[4] oscilleraient autour de 9 :3 :3 :1 mais ne seraient pas toutes exactement de 9 :3 :3 :1. Ainsi, une proportion légèrement différente ne réfute pas l'hypothèse initiale car on sait que, d'un échantillon à l'autre, la répartition peut varier légèrement. Pour mesurer le niveau auquel les observations réfutent l'hypothèse d'indépendance (appelons-la $H_{9:3:3:1}$), Fisher procède de la manière suivante.

2. Primevère

3. Dans la population d'intérêt, c'est à dire l'ensemble hypothétique et de taille infinie comprenant tous les résultats des croisements de *Primula* de génération hybride

4. Au sein d'échantillons de taille finie prélevés dans la population d'intérêt

Il utilise une statistique, χ^2, mesurant le niveau auquel les observations s'éloignent des valeurs attendues sous $H_{9:3:3:1}$. La formule du χ^2 est la suivante :

$$\chi^2 = \sum_{i=1}^{4} \frac{x_i^2}{m_i}$$

Avec x_i : les écarts entre les fréquences attendues et observées dans chacune des quatre cellules du tableau (voir tableau 2.1), m_i les fréquences attendues dans ces quatre mêmes cellules. Une grande valeur de χ^2 correspondra donc des observations qui s'écartent fort de ce qui est attendu sous $H_{9:3:3:1}$ et inversement.

Si la répartition des allèles était bien de 9 :3 :3 :1, alors, les écarts entre les fréquences observées et les fréquences théoriques mesurés par le χ^2 devraient suivre une distribution du χ^2 à trois degrés de liberté (voir figure 2.1). Sur cette distribution on peut voir que, lorsque cette hypothèse est correcte, la plupart du temps les valeurs de χ^2 mesurées se situeront entre 0 et 5, quelques fois entre 5 et 10 et rarement au-delà de 10.

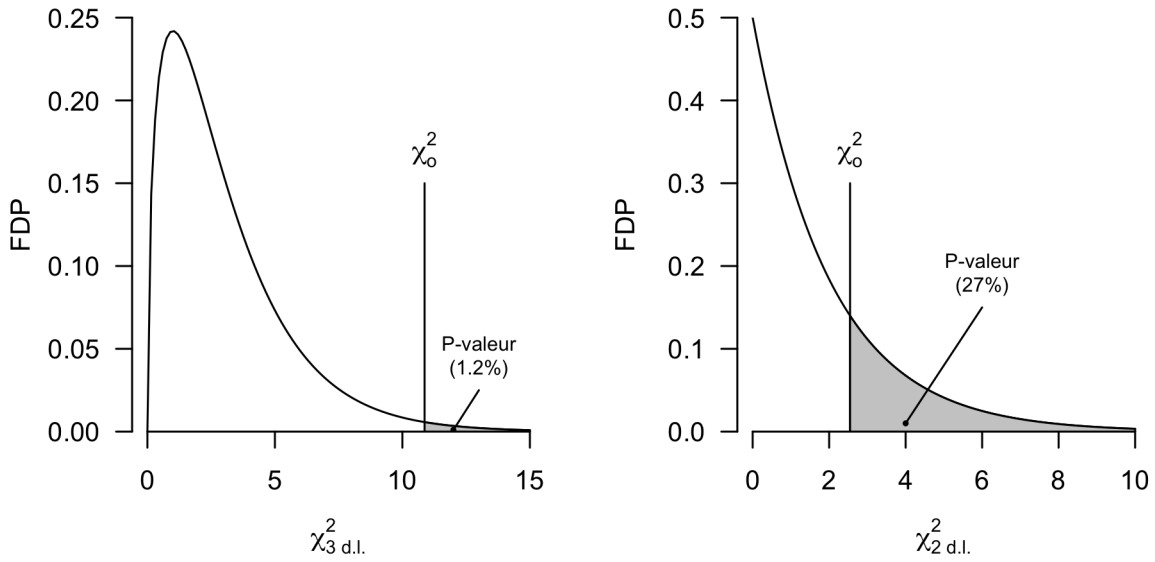


FIGURE 2.1 – **Représentation de la *P*-valeur dans l'exemple de la répartition des allèles de primevères**. Représentation de la distribution théorique du χ^2 à 3 (gauche) et 2 (droite) degrés de libertés. Dans chacun des cas, la valeur de χ^2 observée est représentée (χ^2_o) ainsi que la P-valeur (zone grisée). FDP : fonction de densité de probabilité.

Or, la valeur de χ^2 obtenue avec les observations du tableau 2.1 est de 10.87, soit une valeur plutôt rare sous cette distribution. On peut calculer la probabilité de dépasser cette valeur lorsque $H_{9:3:3:1}$ est correcte ; il s'agit de ce que Fisher appelle *niveau de significativité*, et qui sera plus tard appelé P-valeur.

$$P - valeur = P(\chi^2 \geq 10.87 | H_{9:3:3:1}) = 1.2\ \%$$

Dans cet exemple, les résultats sont plutôt incohérents avec $H_{9:3:3:1}$, car sous l'hypothèse d'une répartition mendeléenne de ces deux traits génétiques, il y a seulement entre 1 % et 2 % de chances d'observer des écarts aussi importants entre les valeurs attendues et observées. Fisher conclut donc que les données ne sont pas cohérentes avec l'hypothèse initiale.

Il continue ensuite son raisonnement en cherchant une explication au fait que les données s'écartent de la répartition 9 :3 :3 :1. Une explication possible serait que les plantes aux feuilles ondulées n'auraient pas la même viabilité que celles aux feuilles plates. Il va alors tester l'hypothèse selon laquelle la répartition de l'aspect de la fleur est bien de 3 :1 au sein de chaque type de feuille (voir tableau 2.2).

TABLE 2.2 – **Valeurs observées et attendues sous l'hypothèse d'une répartition mendeléenne de l'aspect du centre de la fleur au sein de chaque type de feuille**.

	Flat Leaves.		Crimped Leaves.		χ^2.
	Normal Eye.	Primrose Queen Eye.	Lee's Eye.	Primrose Queen Eye.	
Observed . .	328	122	77	33	...
Expected . .	337·5	112·5	82·5	27·5	...
x^2/m . . .	·267	·804	·367	1·109	2·547

Source : [Fisher, 1925]

Cette fois, les résultats ne contredisent pas l'hypothèse énoncée car la P-valeur est d'au moins 20 % (voir figure 2.1). La discordance entre les données observées et l'hypothèse de départ (répartition 9 :3 :3 :1) s'explique donc par une répartition de feuilles différente du 3 :1 qui pourrait s'expliquer par une viabilité moindre des *Primula* aux feuilles ondulées [Fisher, 1925].

Type de tâche

Dans le précédent exemple, on peut voir que, selon Fisher, le test de significativité a pour but de confronter des observations à une hypothèse ou, plus exactement, aux prédictions que l'on pourrait faire sous cette hypothèse.

On pourrait, dès lors, dire que le type de tâche du test de significativité serait : "*Mesurer le degré auquel une série d'observations corrobore une hypothèse probabiliste*"

> *(...) the tests of significance [are a mean] by which we can examine whether or not the data are in harmony with any suggested hypothesis.* [p.8][Fisher, 1925]

Technique

La *technique* que Fisher propose pour réaliser cette te type de tâche peut se décomposer de la manière suivante (d'après [Mayo and Spanos, 2006]).

1. **Énoncer l'hypothèse d'intérêt,** H ;

   Cette hypothèse doit être suffisamment précise pour élaborer un modèle statistique permettant de faire des prédictions et de les comparer aux observations.

   Dans l'exemple précédent, il y a eu deux hypothèses de ce type : H_1 : l'hypothèse selon laquelle la répartition des allèles serait de 9 :3 :3 :1 et H_2 : l'hypothèse selon laquelle la répartition serait de 3 :1 pour chacun des allèles.

2. **Définir** $d(X)$**, la mesure d'écart entre les observations (**O**) et l'hypothèse** ;

   Pour pouvoir confronter H et O, Fisher propose d'utiliser des statistiques, $d(X)$, calculées à partir des données. Le χ^2 est un exemple de cette statistique, mais on pourrait aussi citer la statistique z dans la comparaison d'une série d'observations à un modèle, la statistique t dans la comparaison d'une différence de moyenne à un standard, la statistique F (de Fisher) dans la comparaison de variances, *etc.* Toutes ces statistiques ont comme caractéristiques d'être d'autant plus grandes (en valeur absolue pour le z et le t) que O s'écarte de H.

   Dans son livre, Fisher propose donc des statistiques adaptées à différentes situations, à la confrontation d'observations de différents types (effectifs, moyennes, variances, *etc.*) avec des hypothèses.

3. **Modéliser la distribution attendue de** $d(X)$ **si** H **est vraie** ;

   Une fois la statistique définie, il convient de déterminer sa distribution théorique attendue sous H. Dans le cadre de la comparaison d'effectifs observés à des effectifs théoriques, Fisher nous indique que la distribution du χ^2 à 3 degrés de libertés correspond à ce qui serait attendu sous l'hypothèse H_1 tandis que pour H_2 la distribution attendue serait la distribution du χ^2 à deux degrés de liberté.

   De manière générale, Fisher propose pour chacune des statistiques, une distribution théorique attendue, un modèle statistique. Il ne donne pas de preuve mathématique justifiant l'utilisation de ces distributions théoriques plutôt que d'autres.

4. **Calculer** $d(x_o)$**, l'écart observé entre** H **et** O ;

   A partir d'observations O, d'une hypothèse H et d'une manière de mesurer l'écart entre O et H, Fisher propose de calculer la statistique observée, $d(x_o)$.

   Dans l'exemple des primevères, l'écart entre les observations et H_1 (répartition 9 :3 :3 :1) se mesure avec la statistique observée $\chi^2_o = 10.87$, l'écart entre O et H_2 sera mesuré par le $\chi^2_o = 2.547$.

5. **Calculer la *P*-valeur**, c'est-à-dire la mesure du niveau auquel O corrobore H : $P-valeur = P(d(X) \geq d(x_o)|H)$.

   Dans l'exemple des primevères, la P-valeur mesurant le niveau auquel O corrobore H_1 est de 1.2 % tandis qu'elle est de 27 % entre O et H_2.

A ces étapes, on pourrait ajouter celle de l'**interprétation de la *P*-valeur** qui est loin d'être automatique pour Fisher.

Une P-valeur proche de 0 indique que les observations ne sont pas cohérentes avec H et que celle-ci est réfutée. Une P-valeur élevée indique que les observations ne permettent pas de réfuter H. H est donc renforcée mais n'est jamais validée.

Mais qu'est-ce qu'une P-valeur faible ou élevée ? A première vue, on peut utiliser la convention suivante :

> *"If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Whe shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy".* [Fisher, 1925]

Cependant, selon Fisher, cette convention ne doit pas épargner au chercheur le travail d'interpréter la P-valeur exacte dans son contexte expérimental et théorique.

Pour prendre un exemple, un résultat associé à une P-valeur de 4 % ne remettra pas autant en question l'hypothèse initiale si elle provient d'une analyse statistique dans laquelle on trouve un très grand nombre de tests ou s'il s'agit de la P-valeur associée au critère de jugement principal (*primary endpoint*) dans un essai clinique. La P-valeur ne permet pas de tirer une conclusion en elle-même mais elle doit être intégrée dans un jugement critique qui intègre les différents éléments de contexte autour des observations.

> *"Fisher suggested that it be used as part of the fluid, non-quantifiable process of drawing conclusions from observations, a process that included combining the P value in some unspecified way with background information".* [Goodman, 1999]

Une P-valeur faible indiquera donc que les données s'écartent du modèle et sont *significatives* dans le sens où elles méritent un second examen. Cela ne veut pas encore dire que l'effet étudié est démontré. Pour cela, il faut que le scientifique puisse reproduire cet effet.

> *"The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant re-*

sults which he does not know how to reproduce are left in suspense pending further investigation". [Fisher, 1929]

Technologie

Comment Fisher justifie-t-il cette technique ? Quels éléments de technologie peut-on déceler ? Les éléments de technologie qui viennent justifier cette technique ne sont pas énoncés explicitement. A travers la présentation que Fisher (1925) donne dans son livre *Statistical Methods for Research Workers*, on peut relever les trois éléments suivants :

1. **Justification du choix de la statistique** $d(X)$;
   Fisher justifie le fait que la statistique $d(X)$ (χ^2 dans l'exemple) est une bonne mesure d'écart entre O et H simplement par l'examen de la formule du $\chi^2 = \sum_{i=1}^{n} \frac{x_i^2}{m_i}$ avec m_i : les fréquences attendues sous H dans la classe i et x_i les écarts aux fréquences attendues.

   "*This formula gives the value of* χ^2*, and it is clear that the most closely the observed numbers agree with those expected, the smaller will the* χ^2 *be.*" [Fisher, 1925, p.80]

2. **Justification de la distribution de** $d(X)$ **attendue sous H** ;
   Dans l'exemple qu'il donne, Fisher fait référence à la distribution du χ^2. Mais plutôt que de démontrer en quoi cette distribution théorique est bien celle à laquelle on devrait s'attendre dans ce cas-ci, il renvoie le lecteur au travail de (Karl) Pearson.

   "*For any value of n, which must be a whole number, the form of distribution of* χ^2 *was established by Pearson in 1900 ; it is therefore possible to calculate in what proportion of cases any value of* χ^2 *will be exceeded.* [Fisher, 1925, p.81]

3. **Justification du choix de** P **comme mesure du niveau auquel** O **corrobore** H.
   Si le choix de la P-valeur comme mesure du niveau de corroboration de H peut sembler *a posteriori* évident, il faut rappeler qu'au moins deux autres mesures étaient susceptibles de remplir ce rôle : la statistique $d(x)$ et la vraisemblance $f_H(d(x))$ (voir figure 3).
   La statistique du χ^2 est, en effet, en soi une mesure du niveau auquel les observations corroborent l'hypothèse H puisque celle-ci est d'autant plus grande que les observations s'éloignent de H.
   De même, la vraisemblance $f_H(d(x))$, c'est-à-dire la valeur de la fonction de densité de probabilité associée à la statistique observée, pourrait aussi servir de mesure de corroboration, une grande vraisemblance indiquant un résultat plutôt en accord avec H tandis que cet accord diminue à mesure que la vraisemblance tend vers 0.

   "P [is] therefore the probability that χ^2 *shall exceed any specified value. To every value of* χ^2 *there thus corresponds a certain value of P ; as* χ^2 *is increased from 0 to infinity, P diminishes from 1 to 0".* [Fisher, 1925, p.81]

Fisher préfèrera donc utiliser la P-valeur qui, bien qu'étant directement liée à la statistique $d(x)$, présente l'avantage d'être exprimée sous la forme d'une probabilité, ce qui est plus facile à interpréter que la statistique ou la vraisemblance de la statistique. On pourrait ajouter que la P-valeur aura la même interprétation quelle que soit la distribution théorique sur laquelle elle aura été calculée, ce qui n'est pas le cas de la statistique ou de la vraisemblance.

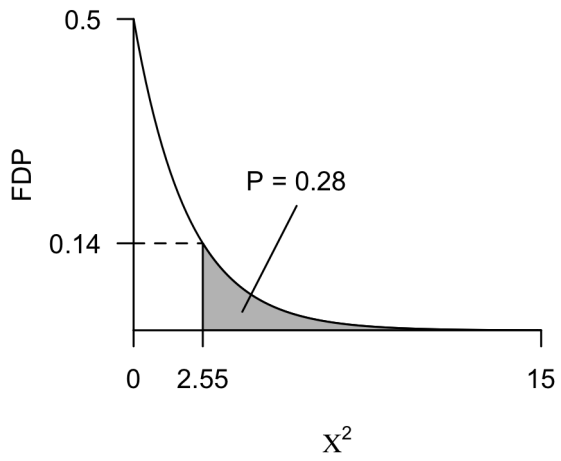


FIGURE 2.2 – **Trois mesures du niveau auquel des observations corroborent une hypothèse**. Ces trois mesures sont : (1) La statistique observée (le χ^2 de 2.55 dans l'exemple), (2) La vraisemblance de cette observation sous H, c'est-à-dire la valeur de la fonction de densité de probabilité pour un χ^2 de 2.55, $f_H(2.55) = 0.14$ et (3) la P-valeur, ici 28 %.

Théorie

Quels éléments théoriques justifiant cette technologie peut-on identifier dans les écrits de Fisher ? Dans son livre à destination des chercheurs, on peut trouver des explications concernant les idées principales qui orientent le raisonnement de Fisher. Nous en avons épinglé deux :

1. **Lien entre la démarche scientifique et le raisonnement statistique**.

   "The statistical examination of a body of data is thus logically similar to the general alternation of inductive and deductive methods throughout the sciences. A hypothesis is conceived and defined with all necessary exactitude ; its logical consequences are ascertained by a deductive argument ; these consequences are compared with the available observations ; if these are completely in accord with the deductions, the hypothesis is justified at least until fresh and more stringent observations are available." [p.8][Fisher, 1925]

Fisher exprime ici sa vision du raisonnement statistique vu comme une démarche hypothético-déductive dans laquelle les hypothèses sont utilisées pour faire des prédictions qui seront confrontées aux observations. La cohérence entre des observations et une hypothèse ne permet jamais de valider cette dernière de manière définitive. Cette vision est très proche de celle que Popper formalisera quelques années plus tard [Popper, 1935].

2. **La place de la probabilité dans l'inférence statistique**.

Le concept de probabilité est, selon Lehman (1992), un concept avec lequel Fisher s'est débattu durant toute sa carrière. Cela s'explique notamment par le fait que ses travaux sont, pour la plupart, antérieurs à la formalisation des axiomes des probabilités par Kolmogorov en 1933. Dans l'introduction de son livre à destination des chercheurs, Fisher définit la probabilité comme un outil déductif permettant de prédire au sein d'une population les fréquences relatives de différents échantillons. Il insiste sur le fait que, selon lui, cet outil ne peut être utilisé pour exprimer un niveau de croyance dans une hypothèse comme le feront les statisticiens bayésiens.

> *"The deduction of inferences respecting samples, from assumptions respecting the populations from which they are drawn, shows us the position in Statistics of the classical Theory of Probability. For a given population we may calculate the probability with which any given sample will occur, and if we can solve the purely mathematical problem presented, we can calculate the probability of occurrence of any given statistic calculated from such a sample.*
>
> (...)
>
> *For many years, extending over a century and a half, attempts were made to extend the domain of the idea of probability to the deduction of inferences respecting populations from assumptions (or observations) respecting samples. Such inferences are usually distinguished under the heading of Inverse Probability, and have at times gained wide acceptance.*
>
> *This is not the place to enter into the subtleties of a prolonged controversy ; it will be sufficient in this general outline of the scope of Statistical Science to reaffirm my personal conviction, which I have sustained elsewhere, that the theory of inverse probability is founded upon an error, and must be wholly rejected. Inferences respecting populations, from which known samples have been drawn, cannot by this method be expressed in terms of probability, except in the trivial case when the population is itself a sample of a super-population the specification of which is known with accuracy.*
>
> (...)
>
> *The rejection of the theory of inverse probability should not be taken to imply that we cannot draw, from knowledge of a sample, inferences respecting the corresponding population. Such a view would entirely deny validity to all experimental*

> *science. What is essential is that the mathematical concept of probability is, in most cases, inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability.* [p.9-11][Fisher, 1925]"

Fisher se distingue de l'utilisation qui est faite du concept de probabilité dans l'inférence bayésienne (avec ce qu'il appelle la probabilité inverse) et justifie l'emploi d'une probabilité, la P-valeur, pour déterminer si *les données sont en harmonie avec une hypothèse.*

> *"The probabilities established by those tests of significance, which we shall later designate by t and z, are, however, entirely distinct from statements of inverse probability, and are free from the objections which apply to these latter. Their interpretation as probability statements respecting populations constitute an application unknown to the classical writers on probability.* [p.11][Fisher, 1925]"

2.3.2 Test d'hypothèses (B)

Dès 1928, Jerzy Neyman et Egon Pearson proposent une autre manière de tester des hypothèses [Neyman and Pearson, 1928]. Ils s'appuient, entre autre, sur les travaux de Fisher concernant le test de significativité et le principe de vraisemblance. A travers leurs apports théoriques, Neyman et Pearson vont apporter au test de significativité de Fisher des bases mathématiques plus solides. Cependant, comme nous allons le voir, si la manière dont ils reformulent le problème leur permettra effectivement de renforcer les bases mathématiques du test de significativité, cela se fera au prix d'un changement fondamental de la logique sous-jacente au test statistique.

Dans ce qui suit, nous allons présenter la praxéologie B, relative au test d'hypothèses, en partant d'extraits de leur publication majeure de 1933, *On the Problem of the Most Efficient Tests of Statistical Hypotheses.*

Type de tâche

Neyman et Pearson commencent par reformuler la tâche du test statistique.

> *"(...) as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth of falsehood of that hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in follo-*

wing which we insure that, in the long run of experience, we shall not be too often wrong."[Neyman and Pearson, 1933]

"Let us now for a moment consider the form in which judgments are made in practical experience. We may accept or we may reject a hypothesis with varying degrees of confidence ; or we may decide to remain in doubt. But whatever conclusion is reached the following position must be recognised. If we reject H0, we may reject it when it is true ; if we accept H0, we may be accepting it when it is false, that is to say, when really some alternative Ht is true. These two sources of error can rarely be eliminated completely ; in some cases it will be more important to avoid the first, in others the second." [Neyman and Pearson, 1933]

Leur objectif assumé n'est donc pas de savoir si une hypothèse est vraie ou fausse mais de *guider notre comportement vis-à-vis de ces hypothèses*. Par rapport au raisonnement de Fisher, on peut noter deux différences majeures.

La première est qu'ils considèrent plusieurs hypothèses en même temps et non pas une seule à la fois comme c'était le cas avec le test de significativité de Fisher. La seconde est de considérer les conclusions possibles en termes de choix qui peut soit être bon (accepter une hypothèse qui se trouve être la bonne) soit mauvais (accepter une hypothèse quand c'est une autre hypothèse qui est, en réalité, correcte).

Une fois cela posé, Neyman et Pearson définissent deux types d'erreurs :

1. L'erreur de premier type (ou erreur α) qui correspond à la probabilité de rejeter l'hypothèse nulle (que l'on notera H_0) quand celle-ci est, en fait, correcte. $1-\alpha$, la probabilité d'accepter l'hypothèse nulle quand elle est vraie, correspond à la confiance.
2. L'erreur du second type (ou erreur β) qui correspond à la probabilité d'accepter l'hypothèse nulle quand celle-ci est, en réalité, fausse. $1-\beta$, la probabilité de rejeter l'hypothèse nulle quand elle est fausse s'appelle la puissance.

Comme on le voit, une des deux hypothèses possède un statut particulier et sert, en quelque sorte, de référence, il s'agit de l'hypothèse nulle, H_0. Pour cette hypothèse, il sera possible de fixer le niveau d'erreur α toléré.

Neyman et Pearson partent de la représentation graphique suivante pour expliquer leur raisonnement : ils considèrent un univers (*sample space*) à deux dimensions, de façon à ce que chaque échantillon puisse être représenté par un point dans un système à deux dimensions (voir figure 2.3). Dans cette représentation, il existe de nombreuses manières de définir une zone de rejet de H_0 qui corresponde à un risque d'erreur α déterminé. Ils tracent quatre zones : $w1$, $w2$,

la zone à droite du segment BC, $w3$ le cercle autour de A2 et $w4$, la zone définie par le triange ODE.

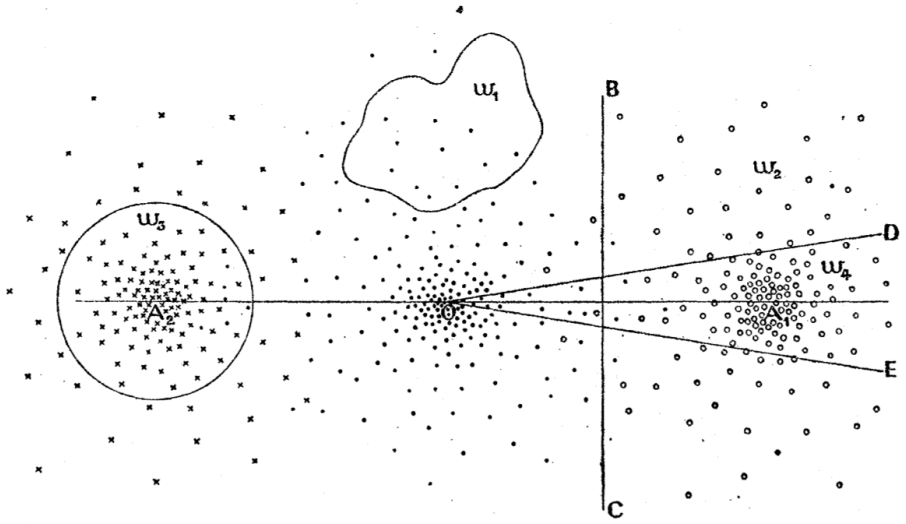


FIG. 1.

FIGURE 2.3 – **Représentation de la distribution des échantillons possibles sous trois hypothèses différentes**. Dans ce système à deux dimensions, les distributions théoriques attendues sous trois hypothèses différentes sont représentées : les points autour du point O correspondent à la distribution attendue sous H_0, les cercles autour de $A1$ correspondent à la distribution attendue sous H_1 et les croix autour de $A2$ correspondent à la distribution attendue sous H_2. Dans cette représentation, la probabilité associée à chaque surface est proportionnelle au nombre de points dans cette surface et la vraisemblance en un point est proportionnelle à la densité de points en cet endroit. Source : [Neyman and Pearson, 1933].

> *"In trying to choose a proper critical region, we notice at once that it is very easy to control errors of the first kind (...). In fact, the chance of rejecting the hypothesis H0 when it is true may be reduced to as low a level as we please. (...) It is when we turn to consider the second source of error - that of accepting H0 when it is false - that we see the importance of distinguishing between different critical regions".* [Neyman and Pearson, 1933]

C'est donc à partir du moment où ils considèrent le risque d'erreur β que Neyman et Pearson parviennent à déterminer quelle région de rejet de H_0 serait la plus intéressante parmi toutes les régions possibles. Ils énoncent que la meilleure région serait la région qui minimise l'erreur β. Ils démontrent par la suite que la meilleure manière de définir cette région est d'utiliser le *rapport de vraisemblance*. Le critère de la minimisation de l'erreur β ou de la maximisation de la puissance sera également déterminant pour définir parmi plusieurs tests statistiques lequel il convient de choisir dans un cas particulier.

A partir de ces éléments, il est possible de formuler le type de tâche du test d'hypothèses : "*Choisir parmi deux hypothèses concurrentes celle qu'il faut considérer correcte tout en contrôlant le risque de première espèce et en maximisant la puissance*".

Technique

Dans leur article, Neyman et Pearson (1933) proposent une technique permettant de déterminer les meilleures régions critiques. Ils donnent l'exemple suivant :

> *Suppose that it is known that a sample of n individuals,* $x_1, x_2, \ldots x_n$ *has been drawn randomly from some normally distributed population with standard deviation* $\sigma = \sigma_0$*, but it is desired to test the hypothesis* H_0 *that the mean in the sampled population is* $a = a_0$*. Then the admissible hypotheses concern the set of populations for which :*
>
> $$p(x) = \frac{1}{\sigma.\sqrt{2\pi}} \times e^{-\frac{(x-a)^2}{2\sigma^2}}$$
>
> *the mean, a, being unspecified but* σ *always equal to* σ_0*. Let* H_1 *relate to the member of this set for which* $a = a_1$*. Let* x *and* s *be the mean and standard deviation of the sample.*
>
> (...)
>
> *Two cases will now arise,*
>
> *(a)* $a_1 < a_0$*, then the region is defined by* $\bar{x} = \frac{1}{n} \times \sum_{i=1}^{n}(x_i) \leq \bar{x_0}$
>
> *(b)* $a_1 > a_0$*, then the region is defined by* $\bar{x} =\geq \bar{x_0}$
>
> (...)
>
> *If, however, the class of admissible alternatives includes both those for which* $a < a_0$ *and* $a > a_0$*, there will not be a single best critical region ; for the first it will be defined by* $\bar{x} \leq \bar{x_0}$ *and for the second by* $\bar{x} \geq \bar{x_0}$*, where* $\bar{x_0}$ *is to be chosen so that* $Po(x \lesseqgtr x_0) = e$*. This situation will not present any difficulty in practice. Suppose* $x > a_0$ *as in fig. 4. We deal first with the class of alternatives for which* $a > a_0$*. If* $e = 0.05$ *;* $X_0 = a_0 1.6449 \frac{\sigma_0}{\sqrt{n}}$*, and if* $x < x_0$*,we shall probably decide to accept the hypothesis* H_0 *as far as this class of alternatives is concerned. That being so, we shall certainly not reject* H_0 *in favour of the class for which* $a < a_0$*, for the risk of rejection when* H_0 *were true would be too great.*

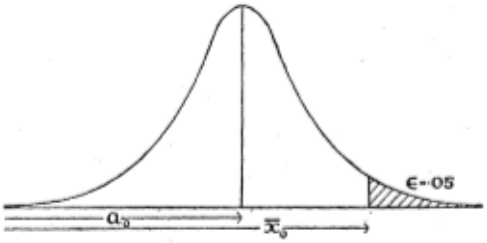


FIG. 4.

> *The test obtained by finding the best critical region is in fact the ordinary test for the significance of a variation in the mean of a sample ; but the method of approach helps to bring out clearly the relation of the two critical regions $x \leq x_0$ and $x \geq x_0$. Further, it has been established that starting from the same information, the test of this hypothesis could not be improved by using any other form of criterion or critical region.* [Neyman and Pearson, 1933]

De cet exemple, on pourrait tirer la technique plus générale suivante, dans le cas où l'on dispose d'un échantillon de n observations :

1. Énoncer deux hypothèses de paramètres définis ;
2. Définir les distributions attendues pour l'échantillon sous chacune des deux hypothèses ;
3. Définir le risque α toléré ;
4. Identifier la meilleure région critique de taille α. Il s'agit de celle qui maximise la puissance $1 - \beta$, cette région est déterminée à partir du rapport de vraisemblance.

Si, initialement, la tâche qu'ils se sont assignée se rapproche de "*Comment tracer les meilleures régions de rejet d'une hypothèse ?*", la réponse apportée en considérant plusieurs hypothèses dans un test statistique et en prêtant attention à la notion de puissance leur a permis de poser les bases d'une théorie des tests statistiques qui aborde de nouvelles questions telles que "*Quelle est la distribution statistique la plus adaptée pour tester une hypothèse particulière ?*" ou encore "*Quelle devrait être la taille d'échantillon nécessaire pour discriminer efficacement deux hypothèses concurrentes ?*" et de manière plus générale : "*Comment choisir parmi deux hypothèses concurrentes celle qu'il faut considérer correcte tout en contrôlant le risque de première espèce et en maximisant la puissance ?*".

Par rapport à cette question, la technique deviendrait :

1. **Définir** H_0 **et** H_1 [5] ;
2. **Fixer** α **et** β, les risques d'erreurs tolérés si, respectivement, H_0 ou H_1 sont vraies ;
3. **Choisir la statistique de test** $d(x)$;
4. **Déterminer la distribution de cette statistique sous chacune des hypothèse** : $f(d(x)|H_0)$ et $f(d(x)|H_1)$;
5. **Choisir la taille d'échantillon** (N) **nécessaire** pour garantir α et β aux niveaux souhaités ;
   - Pour chaque N considéré, identifier la statistique seuil d_s tel que $P(d(x) > d_s|H_0) = \alpha$ et $P(d(x) < d_s|H_1) = \beta$. Ce seuil est déterminé à partir du rapport de vraisemblance $\frac{f(d(x)|H_0)}{f(d(x)|H_1)}$,
   - Tracer la fonction de puissance montrant comment $1 - \beta$ évolue en fonction de N ;

5. H_0 fera référence à l'hypothèse de référence

6. **Récolter les observations et calculer** $d(x_o)$, la statistique observée ;
7. **Selon que la statistique observée soit au-delà ou en-deça du seuil, considérer que** H_0 **est vraie ou que** H_1 **est vraie**.

Technologie

Le test d'hypothèses proposé par Neyman et Pearson repose principalement sur une démonstration mathématique qu'ils formulent dans leur article de 1933 et qui a reçu le nom de Lemme de Neyman et Pearson.

Il s'agit de la démonstration selon laquelle pour tester une hypothèse simple H_0 contre une autre hypothèse simple H_1, la meilleure manière de définir la région critique de taille α consiste à baser le seuil sur le rapport de vraisemblance [Dodge, 2003].

Dans la praxéologie du test d'hypothèses, c'est ce lemme qui légitime l'utilisation du rapport de vraisemblance pour construire les tests statistiques.

Par ailleurs, le cadre posé par Neyman et Pearson permet d'orienter le choix du test statistique puisque le test à privilégier sera celui qui, dans un cas particulier, démontre la plus grande puissance pour départager deux hypothèses.

Ce cadre permet également de justifier le choix d'une taille d'échantillon avant de réaliser les mesures, de manière à atteindre un équilibre acceptable entre nombre d'individus à inclure et risque d'erreurs α et β.

Théorie

En ce qui concerne les éléments de théorie justifiant la technologie proposée, on peut identifier les deux points suivants :

Comme Fisher, Neyman et Pearson considèrent que le concept de probabilité s'applique pour mesurer les fréquences relatives de certains échantillons sous une certaine hypothèse, $P(O|H)$, mais ne devrait pas s'appliquer à la mesure du degré de confiance à accorder à une hypothèse étant donnée une série d'observations, $P(H|O)$.

Ils s'inscrivent donc tout à fait dans le courant de l'inférence statistique fréquentiste.

> *But in general, we are doubtul of the value of attempts to combine measures of the probability of an event if a hypothesis be true, with measures of the a priori probability of that hypothesis. The difficulty seems to vanish in this as in the other cases, if we regard the* λ *surfaces as providing (1) a control by the choice of* ϵ *of the first source of error (the rejection of* H_0 *when true) ; and (2) a good compromise*

> *in the control of the second source of error (the acceptance of H_0 when some H_t is true). The vague a priori grounds on which we are intuitively more confident in some alternative than in others must be taken into account in the final judgment, but cannot be introduced into the test to give a single probability measure.*

Par contre, à l'inverse de Fisher, leur approche s'écarte de la logique de corroboration /réfutation popperienne. Selon Neyman et Pearson, si on teste une hypothèse particulière, c'est qu'il en existe des alternatives. Et s'il existe plusieurs hypothèses, alors le but du test statistique est de faire le bon choix.

> *It is indeed obvious, upon a little consideration, that the mere fact that a particular sample may be expected to occur very rarely in sampling from [a certain population* Π*] would not in itself justify the rejection of the hypothesis that it had been so drawn, if there were no other more probable hypotheses conceivable.* [Neyman and Pearson, 1928]

2.3.3 Probabilité *a posteriori* (C)

Durant la première moitié du 20^e siècle, un autre courant d'inférence statistique se développe. Il trouve ses origines dans les travaux de Bayes (le théorème de Bayes, 18^e siècle), et Laplace (la probabilité inverse, 19^e siècle) et sera formalisé au 20^e siècle par plusieurs auteurs dont Jeffreys en 1939 ([Berger, 2003]).

Ce courant de pensée semble encore plus diversifié que le courant fréquentiste et a, lui aussi, évolué au cours du temps. Dans ce qui suit, nous allons présenter très brièvement une vision de la logique qui sous-tend l'inférence bayésienne afin de montrer les principales différences avec les outils fréquentistes que sont le test de significativité et le test d'hypothèses.

La différence fondamentale entre ce courant bayésien et le courant fréquentiste réside dans l'interprétation du concept de probabilité. Si l'on suit la définition bayésienne donnée par de Finetti, la probabilité peut être vue comme un degré de croyance.

> *"the degree of belief in the occurrence of an event attributed by a given person at a given instant and with a given set of information"* ([De Finetti, 1974] cité dans [Nau, 2001])

Avec l'inférence bayésienne, il devient possible de dire quelque chose à propos de la probabilité des hypothèses elles-mêmes. Dans ce cadre, l'expérience est vue comme une occasion de mettre à jour le niveau de croyance que l'on avait dans une hypothèse. A partir d'une estimation *a priori* de la probabilité d'une hypothèse et d'une série d'observations, il est possible de calculer la probabilité *a posteriori* de cette hypothèse.

TABLE 2.3 – **Table de contingence à l'issue de l'évaluation d'un test diagnostique**.

	Etat		
Résultat du test	Malade	Sain	Total
Positif	231	32	263
Négatif	27	54	81
Total	258	86	344

Source : [Altman and Bland, 1994b].

Ce cadre ouvre la voie à des praxéologies variées. Dans ce qui suit, nous allons tenter d'en exposer deux variantes (C_1 et C_2). Notons que l'idée de cette présentation n'est pas forcément de définir le plus précisément possible C_1 et C_2 mais plutôt de fournir un point de comparaison pour les praxéologies A et B. C'est pour cette raison que nous nous sommes basés sur des sources secondaires plutôt que sur les auteurs originaux.

Contexte (C_1)

Pour évaluer la fiabilité d'un test diagnostique en médecine, on procède généralement de la manière suivante : le test diagnostique est réalisé dans deux groupes de patients, un groupe que l'on sait à l'avance atteint de la pathologie d'intérêt (les malades, M) et un groupe que l'on sait exempt de cette pathologie (les individus sains, S).

Dans les cas les plus simples, le résultat du test diagnostique est binaire, positif ou négatif. La fiabilité du test diagnostique se mesure, dès lors, à travers la sensibilité et la spécificité [Altman and Bland, 1994b]. La première mesure la proportion de résultats positifs chez les individus malades, $P(+|M)$ et la seconde mesure la proportion de résultats négatifs chez les individus sains, $P(-|S)$. Par exemple, sur base de la table de contingence présentée au tableau 2.3, on peut calculer que la sensibilité est de 231/258, soit 90 % et la spécificité est de 54/86, soit 63 %.

Ceci concerne l'évaluation des caractéristiques intrinsèques de ce test diagnostique. Or, ce qui intéresse surtout les médecins dans leur pratique clinique ce sont, non pas les sensibilité et spécificité du test, mais plutôt les valeurs prédictives positives (VPP) et négatives (VPN) [Altman and Bland, 1994a]. En effet, lorsqu'ils font passer un certain test diagnostique à un de leurs patients, ils ne cherchent pas à savoir "*Quelle est la probabilité qu'un individu malade se retrouve positif à ce test diagnostique*", ce que mesure la sensibilité, mais plutôt "*Quelle est la probabilité que ce patient soit réellement malade sachant qu'il est positif à ce test diagnostique*", ce que mesure la VPP.

Type de tâche (C_1)

On peut donc voir ce problème comme un problème de quantification de la probabilité qu'une hypothèse soit vraie : ici, on cherche à déterminer la probabilité que le patient qui vient de passer le test diagnostique soit bien atteint de la maladie suspectée.

Avant d'avoir réalisé le test diagnostique, le médecin averti pourrait avoir une certaine idée de la probabilité [6] que cette hypothèse soit vraie. Par exemple, il pourrait penser que, au vu des caractéristiques du patient, celui-ci aurait une probabilité d'environ 25 % d'être atteint de la maladie suspectée. On dira que la probabilité *a priori* de l'hypothèse "l'individu souffre de la maladie X" est de 25 %. Une fois que l'individu a passé le test diagnostique, le médecin dispose d'informations nouvelles pour actualiser son niveau de croyance dans l'hypothèse, il peut en estimer la probabilité *a posteriori*.

Le type tâche serait donc :"*Mesurer le degré de confiance à accorder à une hypothèse à partir d'une série d'observations et partant d'une certaine confiance a priori*".

Technique (C_1)

Pour répondre à la question précédemment formulée, on peut mettre en œuvre la technique suivante :

1. **Identifier les différentes hypothèses possibles** ;
   Ici deux hypothèses sont possibles pour l'état du patient : malade (M) ou sain (S).
2. **Déterminer les probabilités associées à chaque hypothèse *a priori*** ;
   A priori, compte tenu des caractéristiques du patient, on peut estimer que la probabilité qu'il soit atteint de la pathologie suspectée est de 25 %. $P(M) = 0,25$ et $P(S) = 0,75$.
3. **Définir la statistique,** $d(x)$;
   Dans notre exemple, la statistique qui résumera l'information contenue dans l'observation sera le résultat du test diagnostique : positif (+) ou négatif (-).
4. **Récolter les observations** ;
   Ici, l'observation se résume au résultat du test. Imaginons que ce test soit positif.
5. **Calculer la vraisemblance de chaque hypothèse associée à cette observation,** $f(d(x)|H)$;
   Si le résultat est +, la vraisemblance de l'hypothèse M sera $f(+|M) = Se = 90\%$, tandis que la vraisemblance de l'hypothèse S sera $f(+|S) = 1 - Sp = 37$ %.
6. **Mettre à jour la probabilité associée à chaque hypothèse** :
$$P(H|d(x)) = \frac{P(d(x)|H) \times P(H)}{\sum_{i=1}^{n} P(d(x)|H_i) \times P(H_i)}$$

6. Au sens bayésien

avec n hypothèses.

Dans l'exemple, seules deux hypothèses sont possibles (M ou S). Le calcul de la VPP suite à l'observation d'un test positif sera :

$$
\begin{aligned}
VPP &= P(M|+) \\
&= \frac{P(+|M) \times P(M)}{P(+|M) \times P(M) + P(+|S) \times P(S)} \\
&= \frac{P(M) \times Se}{P(M) \times Se + P(S) \times (1 - Sp)} \\
&= \frac{0.25 \times 0.90}{0.25 \times 0.90 + 0.75 \times 0.63} \\
&= 0.45
\end{aligned}
$$

Dans le cas d'un test diagnostique positif, on peut estimer à 45% la probabilité *a posteriori* que le patient soit réellement malade [Altman and Bland, 1994a].

Technologie (C_1)

Le principal élément de technologie justifiant la technique précédente est le théorème de Bayes, que l'on peut énoncer de la manière suivante :

> *(Bayes, 1763) states that if $q_1, q_2, \ldots, q_n$ are a set of mutually exclusive events, the probability of q_r, conditional on prior information H and on some further event p, varies as the probability of q_r, on H alone times the probability of p given q_r and H, namely* :
>
> $$Pr(q_r|pH) \propto Pr(q_r|H)Pr(p|q_rH)$$
>
> *If $q_1, q_2, \ldots, q_n$ are exhaustive the constant of proportionality is*
>
> $$(\sum_{r=1}^{n} Pr(q_r|H) \times Pr(p|q_rH))^{-1}$$
>
> [Dodge, 2003]

Dans notre exemple, avec M et S, deux hypothèses exhaustives concernant l'état potentiel du patient et l'évènement p qui est le résultat positif $(+)$ du test diagnostique, on obtient :

$$P(M|+) = \frac{P(M \cap +)}{P(+)}$$

$$= \frac{P(M) \times P(+|M)}{P(M) \times P(+|M) + P(S) \times P(+|S)}$$

$$= \frac{P(M) \times Se}{P(M) \times Se + P(S) \times (1 - Sp)}$$

Théorie (C_1)

Les arguments théoriques plus généraux sont donc ceux qui permettent une certaine interprétation des axiomes de Kolmogorov, laissant une place à l'utilisation du **concept de probabilité** au-delà de la définition fréquentiste.

Parallèlement, on trouve aussi dans la démarche bayésienne, un **retour à l'induction** puisque les observations sont continuellement agrégées afin que l'hypothèse soit toujours plus précise.

Dans l'exemple précédent, nous avons cherché à estimer la probabilité que le patient soit malade ou sain au vu du résultat d'un test diagnostique. Il s'agit d'un cas simple car l'état du patient ne peut prendre que deux valeurs distinctes : malade ou sain. Mais en réalité, le raisonnement bayésien s'applique aussi au cas où le paramètre étudié peut prendre une infinité de valeurs.

Contexte (C_2)

Par exemple, on pourrait vouloir décrire la probabilité π de succès d'un certain traitement thérapeutique. Cette probabilité π est inconnue et peut prendre une infinité de valeurs dans l'intervalle $[0, 1]$. Dans ce cas, notre croyance *a priori* concernant la valeur de π prendra la forme d'une distribution de probabilités. On pourra notamment représenter une certaine ignorance *a priori* quant à la valeur π par une distribution uniforme dans l'intervalle $[0, 1]$ (voir figure 2.4, exemple tiré de [Berry, 2006]).

Type de tâche (C_2)

Dans ce contexte la question initiale sera formulée de la manière suivante : " *Comment évolue la distribution de probabilité associée à un paramètre d'une hypothèse suite à l'observation d'une série de données et en partant d'une certaine distribution de probabilité a priori ?* ".

Technique (C_2)

Pour y répondre, on peut utiliser la technique suivante :

1. **Définir une distribution de probabilité *a priori* sur un paramètre inconnu (θ) d'une population,** $f(\theta)$;

   Dans l'exemple cité dans [Berry, 2006], cette distribution *a priori* est une distribution uniforme sur l'intervalle $[0, 1]$. Autrement dit, *a priori* on juge tout aussi probable que le paramètre π se situe entre 0 et 10 % qu'entre 10 et 20 % ou qu'entre 90 et 100 %.

2. **Définir la statistique qui sera utilisée,** $d(x)$;

   Dans l'exemple, le traitement étudié est appliqué à n patients, la statistique résumant les observations sera le nombre de patients (sur les n) chez qui on observe un succès thérapeutique.

3. **Calculer la fonction de vraisemblance,** $f(d(x)|\theta)$;

   A partir d'un résultat observé, par exemple 4 succès et un échec, on peut calculer la fonction de vraisemblance qui associe à chaque valeur du paramètre θ, une vraisemblance.

4. **Mettre à jour la distribution de probabilité sur le paramètre,** $f(\theta|d(x) = f(d(x)|\theta) * f(\theta)$;

   Ensuite, en combinant la distribution de probabilité *a priori* avec la fonction de vraisemblance, on obtient la distribution de probabilité *a posteriori*. Posons, par exemple, que pour les cinq premiers patients traités, nous observons quatre succès et un échec. Dans ce cas, la distribution *a posteriori* aurait l'allure de la courbe bleue. Si on fait le point après les dix premiers patients dont 7 sont des succès et 3 des échecs, on obtiendrait la distribution *a posteriori* représentée par la courbe verte (voir figure 2.4).

De cette manière on répond à la question initiale puisqu'on obtient une distribution de probabilité *a posteriori* pour le paramètre π.

A partir de cette distribution, on peut calculer la probabilité que le paramètre π se situe dans tel intervalle. Ainsi, sur base de la distribution de probabilité représentée par la courbe bleue dans la figure 2.4 [7], on peut déduire que $P(\pi < 50\ \%) = 11\ \%$

7. Une distribution $\beta(5, 2)$.

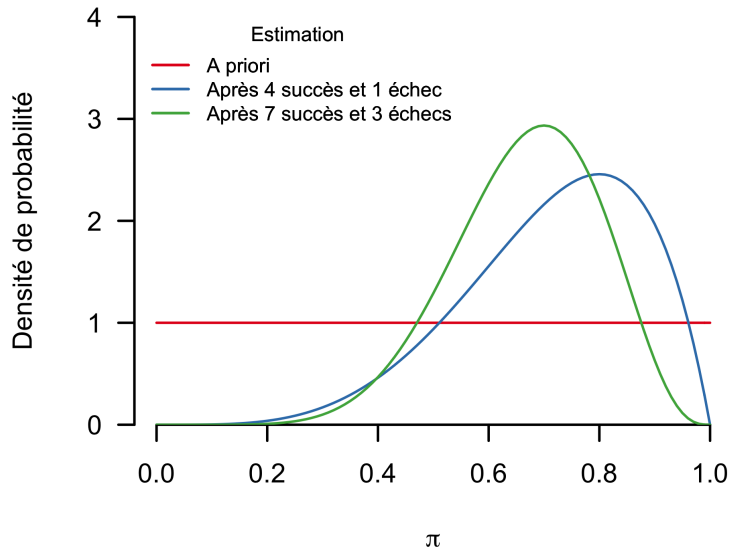


FIGURE 2.4 – **Distributions *a priori* et *a posteriori* dans l'estimation d'un paramètre** π.

Technologie et théorie (C_2)

Les éléments de technologie justifiant la technique et les éléments théoriques justifiant la technologique sont similaires à ceux évoqués dans l'exemple précédent.

2.3.4 Discussion

Après avoir décrit les trois principales praxéologies liées au test statistique qui peuvent se rencontrer dans le domaine du savoir savant, nous allons résumer leurs caractéristiques afin de dégager les points communs et les principales différences que nous illustrerons en les appliquant sur un exemple générique. Nous terminerons par identifier certains points qui posent question pour chacune de ces praxéologies.

Dans cette comparaison, nous simplifierons la praxéologie liée au test d'hypothèses à celle dont la tâche est "Comment choisir parmi deux hypothèses concurrentes celle qu'il faut considérer correcte tout en contrôlant le risque de première espèce et en maximisant la puissance ? (praxéologie B), bien qu'il eût été possible d'en identifier d'autres [8].

De même, concernant la praxéologie liée à la probabilité *a posteriori*, nous nous focaliserons sur le cas simple dans lequel deux hypothèses sont en jeu et où il est possible d'attribuer une probabilité *a priori* à chacune d'elle (C_1).

Le tableau 2.4 résume les trois praxéologies comparées. Voyons, pour chaque niveau praxéologique, en quoi elles diffèrent.

Comparaison des niveaux praxéologiques

Regardons, en premier lieu, **les types de tâches** de ces praxéologie.

Si on voulait trouver un dénominateur commun on obtiendrait une question générale du type :"Comment utiliser des observations pour tirer des conclusions à propos d'hypothèses statistiques [9] ?".

Pour chaque praxéologie cette question générale va se préciser différemment :

— le test de significativité (A) vise à fournir une évaluation quantitative du degré auquel des observations corroborent une hypothèse ;
— le test d'hypothèse (B) pose le problème en termes de choix à opérer parmi plusieurs hypothèses qui peuvent être vraies ou fausses ;
— la probabilité *a posteriori* permet de faire évoluer quantitativement, le degré de confiance qu'un individu accorde à diverses hypothèses à la lumière des nouvelles observations.

Au niveau des **techniques**, on peut noter certains points communs puisque les trois praxéo-

8. Quelle est la meilleure manière de définir les régions critiques ? Comment déterminer le nombre d'individus nécessaires pour départager efficacement deux hypothèses concurrentes ?

9. Par hypothèse statistique, on entendra une *hypothèse qui concerne les paramètres, la forme d'une distribution de probabilité d'une population ou, plus généralement, un mécanisme probabiliste qui est supposé générer les observations* (traduction libre de [Dodge, 2003]).

TABLE 2.4 – Tableau résumé comparant trois praxéologies

Niveau praxéologique	**Test de significativité** (A)	**Test d'hypothèses** (B)	**Probabilité *a posteriori*** (C_1)
Type de tâche	Mesurer le degré auquel une série d'observations corrobore une hypothèse probabiliste	Choisir parmi deux hypothèses concurrentes celle qu'il faut considérer correcte tout en contrôlant le risque de première espèce et en maximisant la puissance	Mesurer le degré de confiance à accorder à une hypothèse à partir d'une série d'observations et partant d'une certaine confiance à priori
Technique	1. Définir H 2. Définir $d(x)$ 3. Modéliser $f(d(x)\|H)$ 4. Calculer $d(x_o)$ 5. Calculer $P(d(x) > d_o\|H)$ 6. Interpréter $P(d(x) > d_o\|H)$	1. Définir H_0 et H_1 2. Fixer α et β 3. Choisir $d(x)$ 4. Modéliser $f(d(x)\|H_0)$ et $f(d(x)\|H_1)$ 5. Calculer N 6. Calculer $d(x_o)$ 7. Si $\frac{f(d(x_o)\|H_0)}{f(d(x_o)\|H_1)} > k$, considérer que H_0 est vraie sinon considérer que H_1 est vraie	1. Définir H_0 et H_1 2. Définir $P(H_0)$ et $P(H_1)$ 3. Choisir $d(x)$ 4. Modéliser $f(d(x)\|H_0)$ et $f(d(x)\|H_1)$ 5. Calculer $P(H_0\|d(x_0))$ et $P(H_1\|d(x_0))$ 6. Interpréter $P(H_0\|d(x))$ et $P(H_1\|d(x))$
Technologie	Intuition, justifications	Lemme de Neyman et Pearson	Théorème de Bayes
Théorie	Démarche de corroboration Probabilité fréquentiste	Démarche de choix Probabilité fréquentiste	Démarche inductive Probabilité bayésienne

logies impliquent, d'une manière ou d'une autre, le test d'une hypothèse statistique.

On trouve donc chaque fois :

1. la définition d'une ou plusieurs hypothèses statistiques (H) ;
2. l'utilisation d'une statistique $(d(x))$ résumant l'information contenue dans l'échantillon ;
3. une distribution de probabilité concernant cette statistique sous chacune des hypothèses considérées $(f(d(x)|H))$.

Une fois les hypothèses statistiques, les statistiques et les distributions de probabilités définies, on peut noter des différences au niveau des informations quantitatives qui entrent en jeu dans les trois techniques.

Dans le cas du test de significativité (A), il n'y a aucune information quantitative supplémentaire qui rentrerait dans le calcul de la P-valeur puisque celle-ci est déduite de ces trois éléments de base. Par contre, Fisher propose que cette P-valeur soit ensuite utilisée dans un *jugement non quantitatif* intégrant les éléments de contexte de l'expérience (plausibilité de l'hypothèse testée, reproductibilité de l'expérience, *etc.*) afin de tirer une conclusion générale à propos de l'hypothèse scientifique en jeu.

A l'inverse, dans le cas du test d'hypothèses (B), ces éléments de contexte (plausibilité initiale des hypothèses, coûts relatifs des erreurs faites en rejetant H_0 ou H_1 à tort, *etc.*) sont censés être utilisés dès le départ au moment où les risques d'erreurs α et β sont définis. L'interprétation du résultat de l'expérience est alors automatique : on considère que H_0 est vraie si la statistique observée est supérieure à un certain seuil et sinon, on considère que H_1 est vraie.

On voit ici que, bien que reposant sur des bases mathématiques similaires, la manière d'utiliser le résultat observé pour tirer une conclusion sur l'hypothèse scientifique générale est tout à fait différente.

La probabilité *a posteriori* (C) implique, elle aussi, d'intégrer de manière quantitative les éléments d'information liés au contexte. Cette intégration se fait au travers du degré de croyance accordé à chaque hypothèse, c'est-à-dire à travers sa probabilité *a priori*.

On peut donc dire, d'une certaine manière, dans la praxéologie A, que l'intégration des résultats au corpus de connaissances est un procédé non quantifiable, qui fait intervenir les capacités de discernement du scientifique, là où dans les deux autres approches, cette intégration se fait de manière quantitative à travers la définition des risques d'erreurs tolérés ou des probabilités *a priori* sur les hypothèses.

Une autre différence entre le test d'hypothèses (B) et le test de significativité (A) concerne le moment auquel les observations sont censées intervenir. Dans la technique du test d'hypothèses, les cinq premières étapes sont nécessairement antérieures aux observations, notamment puisqu'il

s'agit de calculer le nombre d'individus à inclure dans l'échantillon. A l'inverse, avec le test de significativité, les observations peuvent être présentes dès le début de la démarche. En ce sens, on pourrait dire que la démarche du test d'hypothèses est principalement *a priori* (dans le sens où la plus grande partie du raisonnement se fait sans données) tandis que la démarche du test de significativité se fait principalement *a posteriori*.

Le calcul de la probabilité *a posteriori* (C) se fait, comme son nom l'indique, après avoir récolté les observations mais implique, cependant, d'avoir déterminé au préalable le niveau de croyance initial dans chacune des hypothèses.

Si on regarde les ensembles **technologico-théoriques**, on note encore des différences marquantes entre les trois praxéologies.

Premièrement, on peut souligner des différences au niveau du sens donné au concept de probabilité.

Dans la praxéologie du test de significativité (A), la probabilité se mesure dans une distribution théorique, au niveau d'une population hypothétique. Fisher calcule la probabilité d'observer un certain ensemble d'évènement sous une certaine hypothèse statistique.

Dans la praxéologie du test d'hypothèses (B), la probabilité est vue comme une fréquence relative à long terme. Par exemple, en fixant le risque d'erreur α à 5 %, on s'assure que, sur le long terme, on ne rejettera H_1 quand elle est vraie que cinq fois sur cent.

Que ce soit dans le test de significativité ou dans le test d'hypothèses, la probabilité reçoit une définition fréquentiste. Neyman et Pearson ainsi que Fisher rejettent l'idée d'appliquer le concept de probabilité aux hypothèses elles-mêmes.

A l'inverse, dans l'inférence bayésienne, en général, et dans la praxéologie basée sur la probabilité *a posteriori*, en particulier, la probabilité prend un sens plus large que la fréquence relative à long terme et peut être utilisée pour mesurer le niveau de confiance à accorder à une hypothèse.

Deuxièmement, on peut observer différentes conceptions de la manière dont la science doit fonctionner.

La praxéologie du test de significativité (A) est clairement dépendante d'une vision selon laquelle la science avance par la corroboration et la réfutation d'hypothèses. La proximité de la technique proposée avec la démarche scientifique constitue d'ailleurs un des arguments justifiant cette technique.

A l'inverse, dans la praxéologie du test d'hypothèses (B), la manière dont la science doit fonctionner n'est pas clairement établie. On peut cependant trouver, derrière la technique proposée, l'idée selon laquelle la science doit fonctionner en utilisant de manière objective les

observations pour choisir parmi plusieurs hypothèses celle qu'il faut considérer comme correcte.

Enfin, on peut considérer que la praxéologie de la probabilité *a priori* (C) repose sur la conception selon laquelle la science avance par induction. En effet, à mesure que les observations s'accumulent les hypothèses statistiques sont révisées, mises à jour dans un processus continu au sein duquel elles ne sont jamais rejetées ou déconstruites. Les hypothèses se constituent donc de manière progressive, par l'agrégation progressive des observations. Cela diffère de la vision poppérienne de la construction de savoir scientifique selon laquelle les hypothèses sont mises à l'épreuve des observations, sont éventuellement rejetées mais ne sont pas continuellement adaptées aux observations (ce qui les rend irréfutables *de facto*).

Une troisième différence que l'on peut noter concerne le niveau de la justification de la technique.

Dans le cas du test de significativité (A), la technique n'est pas entièrement démontrée, Fisher a l'intuition que cette manière de faire est la bonne mais ne le démontre pas mathématiquement. Les arguments sont donc le bon sens, l'intuition, la proximité de la technique proposée avec la démarche scientifique (au sens de Popper).

Pour le test d'hypothèses (B) et pour la probabilité *a posteriori* (C), la justification repose sur des démonstrations mathématiques, le lemme de Neyman et Pearson pour le premier et le théorème de Bayes pour le second.

On le voit dans cette analyse comparative, les trois praxéologies divergent à bien des niveaux. Cependant, on pourrait se demander quel est l'impact de ces différences dans un certain contexte. Quelles sont les conséquences concrètes de l'application d'une praxéologie plutôt qu'une autre ? Ainsi, afin de donner une idée plus concrète des différences entre les praxéologies, nous allons tenter de les appliquer à un même contexte.

Application à un exemple générique

Dans notre présentation des trois praxéologies, nous avons appliqué le test de significativité à une comparaison d'effectifs avec la statistique du χ^2, le test d'hypothèses à une comparaison d'un échantillon à deux populations à l'aide de la statistique z tandis que la probabilité *a posteriori* a été illustrée sur une distribution binomiale.

Afin d'essayer de rendre les différences plus claires, nous allons tenter d'appliquer chacune des approches sur un même contexte tout en sachant, d'emblée, qu'elles ne répondent pas à la même question initiale.

Le contexte – artificiel – est le suivant : on s'intéresse à une population de patients (par exemple vaccinés contre une certaine maladie) parmi laquelle une certaine proportion π devient immunisée contre la maladie X suite à la vaccination. A propos de cette proportion π, on

évoque généralement les trois hypothèses suivantes :

1. H_1 : 40 % de ces patients sont immunisés ($\pi = 0.4$) ;
2. H_2 : 50 % de ces patients sont immunisés ($\pi = 0.5$) ;
3. H_3 : 90 % de ces patients sont immunisés ($\pi = 0.9$).

On réalise une expérience sur cinq individus que l'on considère tirés aléatoirement dans cette population d'intérêt. Les probabilités de chaque résultat sous chacune de ces trois hypothèses sont représentées à la figure 2.5.

L'objectif sera de "dire quelque chose à propos de ces hypothèses à partir de l'observation d'un échantillon de cinq patients" [10]. Nous allons présenter les conclusions qui sont tirées dans chacune des approches et pour chacun des résultats possibles (0, 1, 2, 3, 4, ou 5 patients immunisés sur les 5).

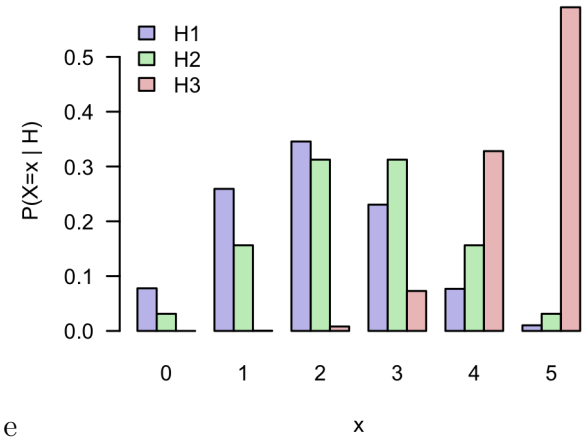


FIGURE 2.5 – **Représentations des probabilités associées à chacun des résultats possibles sous chacune des trois hypothèses**. $H_1 : Bi(5, 0.4)$, $H_2 : Bi(5, 0.5)$ et $H_3 : Bi(5, 0.9)$.

Tentons, tout d'abord, d'appliquer le **test de significativité** (A) à ce contexte.

Comme nous venons de le voir, le test de significativité tente de mesurer le niveau auquel des observations corroborent une hypothèse précise à travers la P-valeur. Dans ce cas on pourrait donc calculer la P-valeur associée à chaque résultat sous chacune des hypothèses (voir tableau 2.5).

On y voit que si notre hypothèse d'intérêt est H_1, alors le résultat $x = 5$ conduira à la remettre en doute (puisque la P-valeur serait de 1 %) ainsi que, dans une moindre mesure, le résultat $x = 4$ (P-valeur = 9 %). Si notre hypothèse d'intérêt est H_2, alors les résultats les plus défavorables seraient $x = 0$ ou $x = 5$ mais ceux-ci restent encore probables sous H_2 (P-valeur =

10. Cet objectif est intentionnellement formulé de manière suffisamment vague pour pouvoir être traité par les trois approches.

TABLE 2.5 – ***P*-valeur associée à chaque résultat sous chacune des trois hypothèses**. La P-valeur a été calculée à partir de la distribution binomiale.

Hypothèse	H_1	H_2	H_3
$x = 0$	0.1648	0.0625	0.0000
$x = 1$	0.6544	0.3750	0.0005
$x = 2$	1.0000	1.0000	0.0086
$x = 3$	0.3952	1.0000	0.0815
$x = 4$	0.0870	0.3750	0.4095
$x = 5$	0.0102	0.0625	1.0000

6 %). Avec si peu d'observations ($n = 5$), aucun résultat ne permettrait une remise en question de H_2. Enfin, si notre hypothèse est H_3, observer $x = 0$, $x = 1$ ou $x = 2$ permettrait de la réfuter avec des P-valeurs respectivement < 0.01 %, < 0.05 % et < 0.9 %.

Appliquons, ensuite, le **test d'hypothèse** tel que nous l'avons défini plus haut (praxéologie B).

Pour "essayer de dire quelque chose" à propos des hypothèses H_1, H_2 et H_3 sur base d'un échantillon de cinq observations, nous pourrions suivre la démarche du test d'hypothèses.

Cela nécessiterait :
— de définir une hypothèse de référence ;
— de fixer le risque α toléré ;
— de définir une hypothèse alternative ;
— de déduire la région d'acceptation de l'hypothèse de référence.

Par exemple, en choisissant H_3 comme hypothèse de référence, H_2 comme alternative et en fixant le risque α à 10 %, on pourrait calculer le risque d'erreur β obtenu. Dans cet exemple, il serait de 18 %, soit une puissance de 82 % (voir figure 2.6, bas droite). Autrement dit, si H_3 était vraie, en la rejetant à partir des observations $x = 0$, $x = 1$, $x = 2$ ou $x = 3$, on ne se tromperait que dans 10 % des cas. De même, si c'était en réalité H_2 qui était vraie, en utilisant les mêmes seuils, on ne se tromperait que dans 18 % des cas. Dans la comparaison H_3 (référence) vs H_2, cette manière de fixer les seuils et de tirer une conclusion aboutirait à une confiance $(1 - \alpha)$ de 90 % et une puissance $(1 - \beta)$ de 82 %.

On peut remarquer qu'avec trois hypothèses au départ et en testant une hypothèse de référence contre une hypothèse alternative, il y a 6 tests d'hypothèses possibles (voir figure 2.6). Pour une même hypothèse de référence, H_2, les résultats qui conduisent à l'accepter ne sont pas les mêmes selon que l'alternative est H_1 ou H_3. De plus, les limites d'acceptation ne sont

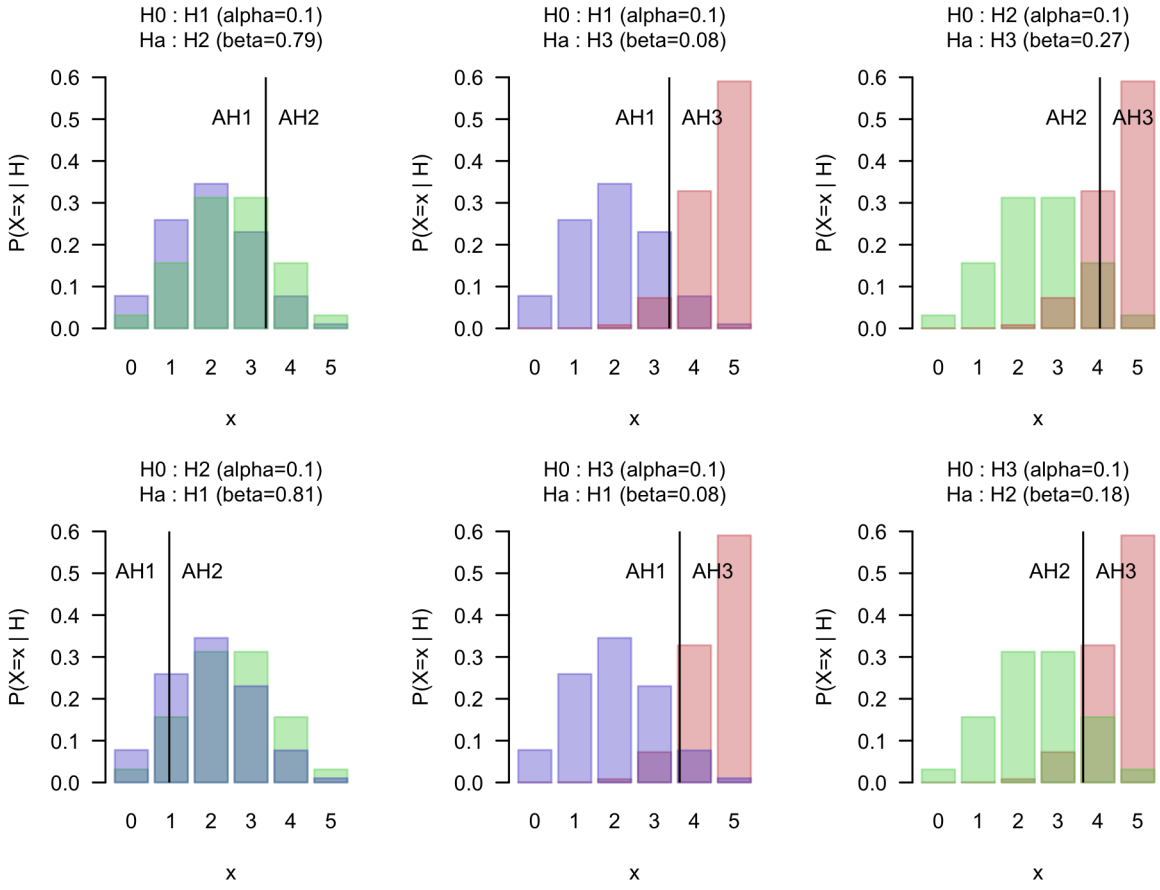


FIGURE 2.6 – **Représentations des probabilités associées à chacun des résultats possibles sous les deux hypothèses dans chacun des tests d'hypothèses possibles**. H_1 : $Bi(5, 0.4)$, H_2 : $Bi(5, 0.5)$ et H_3 : $Bi(5, 0.9)$. Pour chaque test d'hypothèses, le risque α a été fixé arbitrairement à 10 % et le risque β est donné. Le seuil séparant les zones d'acceptation de l'hypothèse nulle et d'acceptation de l'hypothèse alternative est représenté par le trait vertical.

pas les mêmes selon que l'hypothèse de référence soit H_1 ou H_2.

Un autre élément remarquable est le fait que le seuil entre les régions d'acceptation et de rejet de H_0 tombe parfois au milieu d'une valeur (par exemple, sur le graphique en haut à droite de la figure 2.6). Dans ce cas, une même valeur, $x = 4$ en l'occurence, peut conduire à accepter H_2 ou H_3. Intuitivement, on pourrait vouloir la placer juste avant $x = 4$ ou juste après mais pas au milieu de $x = 4$. Et pourtant, si on suit la logique de Neyman et Pearson, placer la limite de la sorte peut avoir du sens. En effet, si on accepte H_3 dans 100 % des cas où l'on observe $x = 4$, cela conduirait à une erreur α de 19 %, ce qui est bien plus que le 10 % voulu, et une erreur β de 8 %. D'autre part, si on accepte H_2 dans 100 % des cas où l'on observerait $x = 4$, cela conduirait à une erreur α de 3 % (ce qui est en dessous des 10 % tolérés) mais une erreur β qui monte à 41 %. Bref, une solution envisageable et qui permet de maximiser la puissance du test tout en conservant le risque α à un niveau prédéfini consiste, lorsque le résultat $x = 4$ est obtenu, à tirer au sort l'hypothèse acceptée entre H_2 et H_3 [Christensen, 2005]. Ici, en donnant

une probabilité de 56 % pour H_2 et 44 % pour H_3 dans ce tirage au sort, on obtiendrait un test d'hypothèses ayant un risque α fixé à 10 % et une puissance maximale.

Essayons maintenant d'appliquer la praxéologie C_1 basée sur la **probabilité *a posteriori***.

Pour appliquer l'approche bayésienne à l'exemple générique, il serait nécessaire de définir le niveau de croyance que l'on a *a priori* dans les hypothèses H_1, H_2 et H_3. Il apparait assez facilement qu'il existe une infinité de solutions possibles pour ce choix. Nous allons considérer deux cas de figure différents : dans le premier, toutes les hypothèses sont équiprobables, elles ont toutes une chance sur trois d'être correctes, tandis que dans le second l'hypothèse H_3 sera considérée comme nettement plus probable que les deux autres. Dans ce cas, les probabilités *a priori* associées aux hypothèses H_1, H_2 et H_3 sont respectivement de 10 %, 10 % et 80 %.

A partir de ces probabilités *a priori* et selon le résultat de l'observation (nombre de patients immunisés sur l'échantillon de 5 patients), on obtiendrait les probabililités *a posteriori* suivantes (voir tableau 2.6).

TABLE 2.6 – **Probabilités *a priori* et *a posteriori* des hypothèses H_1, H_2 et H_3 dans les 2 cas de figure**.

	Cas 1			**Cas 2**		
Hypothèse	H_1	H_2	H_3	H_1	H_2	H_3
	Probabilités *a priori*			Probabilités *a priori*		
	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{8}{10}$
Résultat	Probabilités *a posteriori*			Probabilités *a posteriori*		
$x = 0$	0.7133	0.2866	0.0001	0.7128	0.2865	0.0007
$x = 1$	0.6232	0.3757	0.0011	0.6185	0.3729	0.0086
$x = 2$	0.5188	0.4691	0.0122	0.4781	0.4323	0.0896
$x = 3$	0.3741	0.5075	0.1184	0.2046	0.2775	0.5179
$x = 4$	0.1369	0.2785	0.5847	0.0269	0.0547	0.9184
$x = 5$	0.0162	0.0494	0.9343	0.0021	0.0066	0.9913

On y voit que, dans le premier cas de figure (hypothèses équiprobables au départ), si on observe un patient immunisé sur les cinq ($x = 1$), alors on peut conclure que H_1 a 62.3 % de chances d'être vraie, H_2 a 37.6 % et H_3 0.1 %. Si on observe $x = 3$, alors les probabilités sont respectivement de 37.4 %, 50.8 % et 11.8 %. Autrement dit, après avoir observé un échantillon de cinq patients dont trois sont immunisés, la probabilité que, au niveau de la population, 90 % des patients soient immunisés est passée de 33.3 % à 11.8 %. Par contre, dans le deuxième cas de figure (H_3 beaucoup plus probable *a priori* que H_1 et H_2), observer ce même échantillon aurait fait passer la probabilité de H_3 de 80.0 % à 51.8 % (voir tableau 2.6).

Appliquons, enfin, la praxéologie C_2, également basée sur la **probabilité *a posteriori***, à ce contexte générique.

Nous pouvons, en effet, traiter le problème en considérant que les valeurs que peuvent prendre π, la proportion de patients immunisés dans la population, sont continues dans l'intervalle $[0, 1]$ [11].

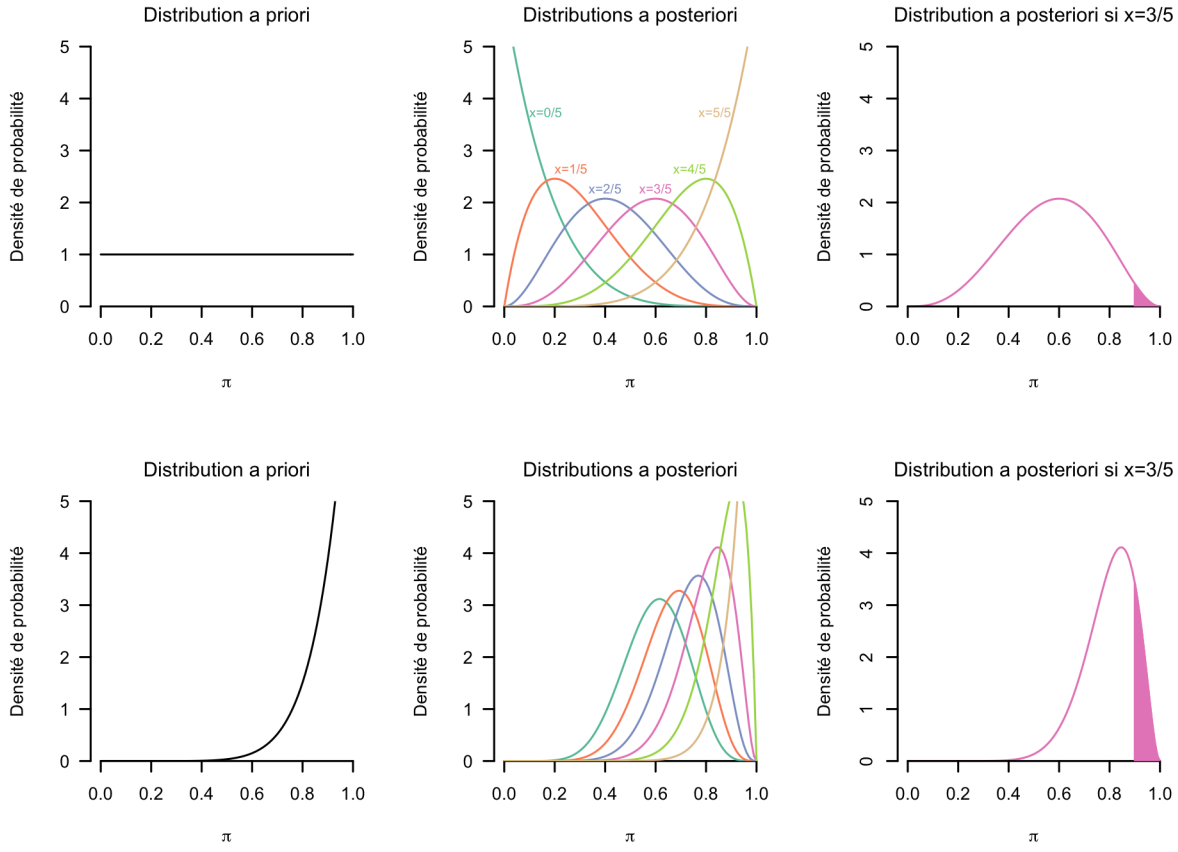


FIGURE 2.7 – **Distributions *a priori* et *a posteriori* dans deux cas de figure**. Haut gauche : distribution *a priori* uniforme. Haut milieu : représentation de toutes les distributions *a posteriori* possibles en fonction du résultat et en se basant sur la distribution *a priori* tracée en haut à gauche. Haut droite : représentation de la distribution *a posteriori* obtenue à partir de la distribution *a priori* uniforme et de l'observation d'une proportion de succès de $3/5$. Bas gauche : distribution *a priori* non uniforme, de loi $\beta(9, 1)$. Bas milieu : représentation de toutes les distributions *a posteriori* possibles en fonction du résultat et en se basant sur la distribution *a priori* tracée en bas à gauche. Bas droite : représentation de la distribution *a posteriori* obtenue à partir de la distribution *a priori* de loi $\beta(9, 1)$ et de l'observation d'une proportion de succès de $3/5$. La surface sous la courbe rose correspond à la probabilité *a posteriori* que $\pi > 0.9$.

11. Ce qui, dans le cas présent, aurait sans doute plus de sens que de se positionner par rapport à trois hypothèses précises.

Dans ce cas, il aurait fallu représenter notre niveau de croyance *a priori* par une distribution de probabilités concernant la valeur du paramètre π. Par exemple, nous aurions pu représenter une certaine ignorance dans la valeur que peut prendre π par une distribution uniforme sur $[0, 1]$ (voir figure 2.7, haut gauche). Ensuite, selon le résultat obtenu au cours de l'expérience, nous aurions adapté la distribution de probabilité (voir figure 2.7). Par exemple, l'observation de trois patients immunisés sur les cinq nous conduirait à la distribution de probabilité rose à partir de laquelle on pourrait déduire qu'il y a seulement 1.6 % de chances que la vraie valeur de π soit d'au moins 90 % (la surface sous la courbe entre $\pi = 0.9$ et $\pi = 1.0$).

Cependant, on aurait pu choisir une autre distribution de probabilités *a priori*, par exemple une distribution reflétant notre croyance dans l'idée que la grande majorité des patients sont immunisés suite à la vaccination (voir figure 2.7, bas gauche), ce qui donnerait lieu à d'autres distributions *a posteriori* (voir figure 2.7, bas milieu). Dans ce cas, l'observation de trois patients immunisés sur cinq patients nous amènerait à conclure qu'il y a 15.8 % de chances que $\pi > 0.9$. On pourrait, de la même manière, calculer les probabilités *a posteriori* que $\pi < 0.4$ ou que $0.5 < \pi < 0.9$, *etc.*

En résumé, voici quelques conclusions auxquelles on aboutirait sous chacune des praxéologies à partir de l'observation de 3 patients immunisés sur les 5 observés.

— A : les observations sont cohérentes avec H_1 et H_2 et un peu moins avec H_3 (pour laquelle la P-valeur vaut 8 %) ;
— B (comparaison H_1/H_2) : un échantillon de 5 patients ne permet pas de discriminer efficacement H_1 de H_2 (erreur β proche de 80 %, erreur α fixée à 10 %) ;
— B (comparaison H_1/H_3) : un échantillon de 5 patients permet de discriminer ces deux hypothèses ;
— C_1 (Cas 1) : l'hypothèse la plus probable *a posteriori* est H_2 ;
— C_1 (Cas 2) : l'hypothèse la plus probable *a posteriori* est H_3 ;
— C_2 (Cas 1) : le paramètre π a 95 % de chances de se trouver entre 22 % et 88 % ;
— C_2 (Cas 2) : le paramètre π a 95 % de chances de se trouver entre 57 % et 95 %.

Limites

Pour achever cette comparaison, nous proposons de passer en revue les principales limites inhérentes à ces trois praxéologies.

Comme on a pu le constater, le principe qui sous-tend **le test de significativité** (A) est très proche de l'idée de corroboration/réfutation développée par Popper. Cette méthode devrait donc *a priori* assez bien convenir au scientifique qui souhaite confronter des observations à une hypothèse en particulier.

Toutefois, nous pouvons identifier plusieurs critiques à ce raisonnement.

Premièrement, on peut se demander comment définir exactement les "données qui s'écartent au moins autant de H que les données observées". Sur la distribution du χ^2 (voir figure 2.1), il est assez aisé d'identifier les valeurs qui correspondent à des observations plus éloignées de H que la valeur observée. Mais qu'en est-il avec d'autres distributions ?

Par exemple, avec une distribution normale, des valeurs très grandes ou très petites correspondent toutes les deux à un écart important à H. Dans ce cas, on dira que les valeurs qui réfutent au moins autant H sont les valeurs dont la fonction de densité de probabilité (FDP) est au moins aussi faible que la valeur observée (voir figure 2.8).

De même, avec la distribution binomiale, comment définir les résultats qui seraient moins en accord avec l'hypothèse que le résultat observé ? Prenons l'hypothèse selon laquelle, au niveau de la population, la probabilité de succès est de 30 % et comparons-la à une expérience fictive dans laquelle on aurait observé 4 succès en 10 tentatives. Au vu des probabilités associées à chaque résultat possible sous H, on aurait tendance à affirmer qu'observer 2 succès remet moins en question H que l'observation de 4 succès (voir figure 2.8). Ce faisant, et comme pour l'exemple de la distribution normale, on définit à nouveau les valeurs les plus éloignées de H comme étant les valeurs les moins probables sous H.

Le problème est que, dans le test basé sur la distribution du χ^2 présenté précédemment, le critère utilisé n'était pas celui des valeurs associées à la plus petite fonction de densité de probabilité. Il semble ne pas y avoir de critère unique permettant d'établir ce que l'on entend par "résultats qui réfutent au moins autant H que les résultats observés".

Deuxièmement, on pourrait remettre en question l'utilisation de la P-valeur comme mesure du niveau auquel des observations réfutent une hypothèse. En effet, celle-ci correspond à la probabilité d'observer ce que l'on a observé ou une valeur plus extrême que celle observée sous l'hypothèse d'intérêt. Mais, dans le fond, pourquoi se baser sur les valeurs plus extrêmes que celle observée si celles-ci n'ont pas été observées ? Selon Jeffreys [12] cela n'a pas de sens car :

> "*a hypothesis that may be true may be rejected because it has not predicted observable results that have not occured*" ([Jeffreys, 1961] cité dans [Berger, 2003]).

Troisièmement, à côté du critère définissant les résultats extrêmes, on ne sait pas encore, au vu des principes évoqués plus haut, sur base de quel critère il faut préférer un modèle de distribution théorique à un autre. Lorsque deux distributions théoriques différentes peuvent être appliquées à une même situation, sur quelle base doit-on opérer notre choix ? Sur base de quel critère faut-il comparer différents tests statistiques ?

Grâce à leurs travaux, Neyman et Pearson fournissent un critère clair permettant de com-

12. Un fondateur du courant bayésien

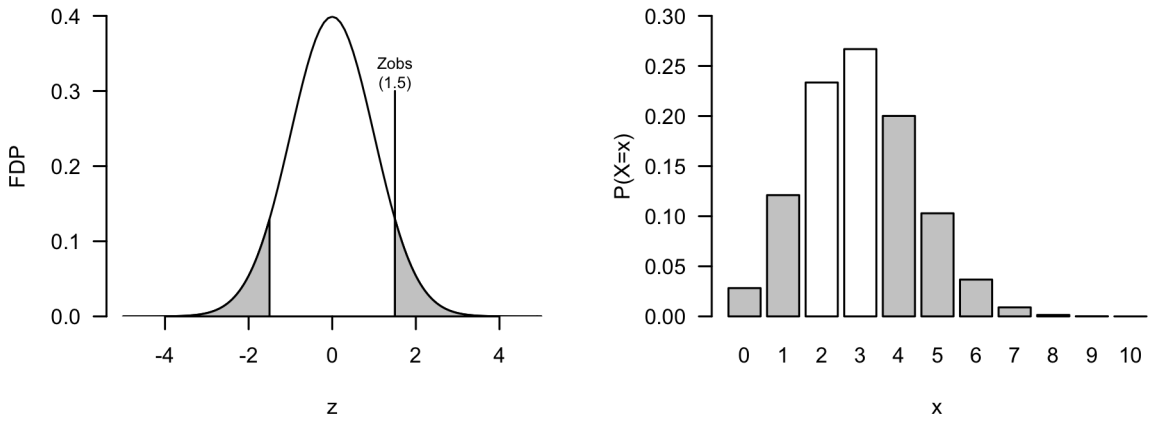


FIGURE 2.8 – **Exemple de distributions binomiale et normale**. Gauche : distribution normale centrée réduite, Z. La zone grisée reprend la P-valeur, c'est-à-dire la probabilité associée à l'ensemble des résultats au moins aussi éloignés du modèle que le résultat observé (z=1.5). Droite : distribution binomiale de paramètre $\pi = 0.3$ et $n = 10$. La surface grisée reprend la P-valeur, c'est-à-dire la probabilité associée à l'ensemble des résultats au moins aussi rares dans ce modèle que le résultat observé ($x = 4$).

parer les tests statistiques entre eux, la puissance, et déduisent que le rapport de vraisemblance permet de définir les meilleures régions d'acceptation et de rejet de l'hypothèse nulle pour un test.

Cela est rendu possible par le fait qu'ils considèrent le problème du test d'hypothèses non pas comme un outil de mesure de la discordance entre une hypothèse spécifique et des observations mais comme un outil permettant de faire un choix objectif entre plusieurs hypothèses concurrentes à partir des observations et en contrôlant les risques d'erreur α et β.

Cependant, en suivant cette logique à la lettre il est possible d'obtenir certains résultats plutôt déconcertants. Par exemple, le test de l'hypothèse A contre l'hypothèse B n'aura pas les mêmes limites et donc les mêmes caractéristiques que le test de l'hypothèse B contre l'hypothèse A. Un autre élément troublant est la possibilité de faire intervenir le hasard pour décider de l'acceptation de l'hypothèse A ou de l'hypothèse B dans certains cas discrets (voir exemple type ci-dessus). Ces éléments sont tout à fait justifiés dans la logique de choix développée par Neyman et Pearson mais n'ont pas de sens dans une logique de mesure du degré d'évidence apporté par les observations en faveur ou contre une hypothèse.

Enfin, on peut trouver troublant l'idée de considérer qu'une hypothèse puisse être vraie ou fausse car on pourrait assez facilement démontrer que toutes les hypothèses statistiques sont fausses dans la mesure où elles ne sont qu'une modélisation d'une distribution au niveau d'une population hypothétique, abstraite. C'est en substance le message derrière l'affirmation "*All models are wrong but some are usefull*" de Box.

> *"For such a model there is no need to ask the question 'Is the model true ?'. If 'truth' is to be the 'whole truth' the answer must be 'No'. The only question of interest is 'Is the model illuminating and useful ?' "* [Box, 1979]

A cause du passage à une logique de décision, les travaux de Neyman et Pearson seront fortement critiqués. D'une part, par Fisher qui considère qu'ils vont trop loin dans cette logique et que celle-ci n'a rien à faire dans la démarche scientifique [Fisher, 1955] et d'autre part, par les tenants de l'inférence bayésienne qui considèrent qu'ils ne vont pas assez loin dans leur logique de décision et qu'il existe d'autres outils inférentiels beaucoup mieux adaptés à cette tâche [Christensen, 2005].

A partir d'une application plus large des axiomes des probabilités et en acceptant l'idée que les savoirs scientifiques se construisent par induction, les tenants de l'inférence bayésienne ont ouvert la voie à tout un champ de l'inférence statistique que nous n'avons que survolé.

Les questionnements soulevés par cette manière de faire de l'inférence statistique ne sont pas d'ordre mathématique mais plutôt philosophique : "Est-il correct d'appliquer le concept de probabilité à un degré de croyance ?", "Peut-on considérer que la science avance par induction ?", "Comment transforme-t-on un certain niveau de confiance dans une hypothèse en une valeur de probabilité ?"

Sous réserve que l'on accepte l'idée d'appliquer le concept de probabilité aux hypothèses, que l'on accepte que la connaissance se fonde, au moins partiellement, sur l'induction et que l'on se donne des règles pour définir les probabilités *a priori*, l'inférence bayésienne dans laquelle s'intègre le concept de probabilité *a posteriori* se révèle être un outil puissant qui a toute sa place dans l'arsenal statistique du scientifique.

2.4 Savoir enseigné

Dans la section précédente nous avons décrit trois praxéologies différentes liées au thème "test statistique" et que l'on rencontre dans ce qu'on pourrait appeler le savoir savant.

Nous allons maintenant tenter de caractériser le savoir enseigné et de définir la praxéologie à laquelle il correspond (praxéologie D).

Nous la comparerons ensuite aux praxéologies A, B et C afin de décrire la transposition didactique et, enfin, nous rechercherons les contraintes institutionnelles qui expliquent que D va nécessairement avoir tendance à s'écarter de A, B ou C.

2.4.1 Démarche enseignée

Dans ce travail, nous nous intéresserons au savoir enseigné aux futurs scientifiques de l'Université de Namur dans l'institution "Cours d'introduction à la biostatistique" . Celui-ci s'appuie sur un dispositif qui peut faire intervenir les éléments suivants : cours *ex cathedra*, capsules vidéos, séances de travaux pratiques, séances de classes inversées, syllabus, site d'auto-apprentissage, site d'évaluation formative et évaluation continue. Certains de ces éléments ont subi de nombreux changements ces dix dernières années (cours *ex cathedra* remplacé par des séances de classes inversées, séances de travaux pratiques souvent remodelées, etc.) tandis que d'autres comme le site d'auto-apprentissage et le syllabus, sont restés relativement stables.

C'est donc à partir de ces sources que nous proposons de décrire le savoir enseigné.

La séquence des thèmes statistiques est la même sur le site et dans le syllabus et a été pensée pour être progressive et modulaire. Ainsi, pour atteindre l'objectif "comparaison de moyennes à l'aide d'un test d'hypothèses", l'étudiant est amené à suivre les modules suivants : analyse descriptive, probabilités, distributions théoriques, distribution normale, principe du test d'hypothèses, gestion des risques d'erreurs, test d'hypothèses relatifs aux variances et enfin, test d'hypothèses relatif aux moyennes (voir figure 2.9).

Par contre, les manières d'expliquer le test d'hypothèses ne sont pas exactement identiques à travers ces deux supports, c'est pourquoi nous les présenterons séparément.

Le syllabus

Comme on peut le voir dans l'extrait suivant, le test d'hypothèses est présenté dans le syllabus comme **un outil de prise de décision concernant l'existence ou non d'un effet**, d'une **information qu'il convient de détecter à travers le bruit**.

> *Le test d'hypothèse est un instrument indispensable en sciences expérimentales, qui permet de rechercher, dans des données x_i affectées d'une variabilité, de l'information partiellement masquée par le bruit dû à ces effets du hasard : Variabilité = information + bruit.*
>
> *Il permet de prendre une décision quant à l'existence de cette information en maîtrisant les risques d'erreur inhérents à cette prise de décision.*
>
> (...)
>
> *Les composants de cet instrument sont les suivants* :
>
> — *<u>un modèle</u> : permettant de décrire le comportement d'une statistique calculée sur les observations sous l'hypothèse que leur variabilité est uniquement due au hasard : variabilité = bruit ; ce modèle et la statistique utilisée dépendent de la nature des données recueillies ;*

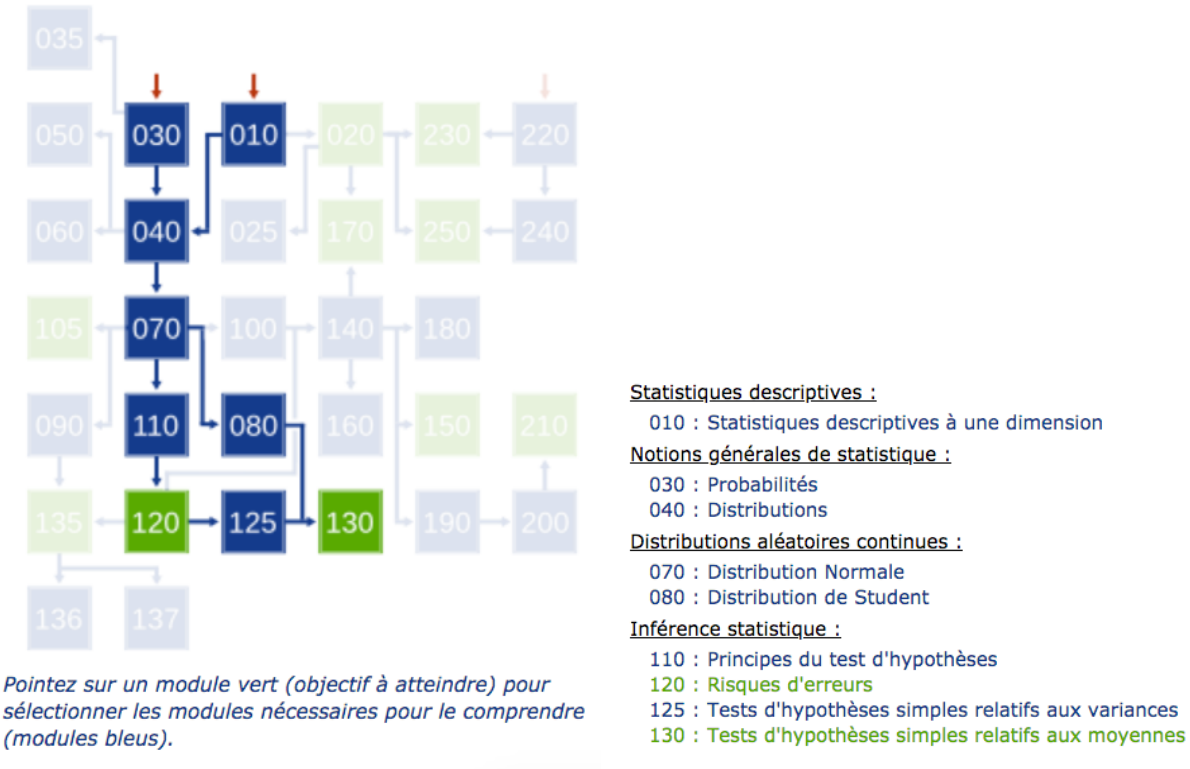


FIGURE 2.9 – **Organisation modulaire des thèmes abordés sur le site d'auto-apprentissage et dans le syllabus**. Ce réseau sémantique représente les modules objectifs (vert) ainsi que les modules qui servent d'appui à ces objectifs. En foncé, on retrouve la séquence de modules à suivre pour atteindre l'objectif "Tests d'hypothèses simples relatifs aux moyennes" [Calmant et al., 2017].

- *une hypothèse nulle (H_0)* : *variabilité = bruit ; nulle signifie : "information non évidente". H_0 permet de donner une (des) valeur(s) de paramètre(s) précis au modèle ;*
- *une hypothèse alternative (H_1)* : *variabilité = information + bruit ; (...)* ;
- *un seuil x_α séparant les valeurs considérées comme probables (de probabilité 1-α) et comme improbables (de probabilité α) dans le modèle H_0 selon l'une des variables modélisant le hasard (Z,t, χ^2, F...) et α est une petite probabilité, de valeur arbitraire souvent fixée à 5% et le seuil est la valeur de la statistique correspondant à cette probabilité dans le modèle. (...)*
- *un échantillon d'observations, qui produit les données sur lesquelles on peut effectuer le calcul de la statistique ;*
- *le calcul d'une statistique ad hoc* ;
- ***la prise de décision*** : *généralement basée sur une p-value, probabilité que les données s'accordent au modèle H_0 ; cette p-value est comparée à la probabilité α : acceptation de l'hypothèse nulle si les observations sont probables suivant le modèle H_0 (statistique en deça du seuil et p-value $\geq \alpha$) ou rejet du modèle H_0 (statistique au delà du seuil et p-value $< \alpha$).*

[Depiereux, 2016, p.156]

Les données d'un échantillon sont résumées à l'aide d'une statistique dont on peut modéliser la distribution attendue sous H_0 (mais pas sous H_1 qui n'est pas entièrement définie). Dans cette distribution, il est possible de définir une zone d'acceptation de H_0 (AH_0) et une zone de rejet de H_0 en fonction du risque d'erreur α que l'on tolère et de la direction du test.

Le test d'hypothèses peut aboutir à deux types de conclusions :

- AH_0, on n'a pas assez d'informations pour rejeter H_0, les données ne s'écartent pas assez du modèle construit sous H_0, l'effet étudié n'a pas pu être mis en évidence ;
- RH_0, l'effet étudié a été mis en évidence, les données s'écartent suffisamment du modèle H_0 pour le rejeter.

Le chercheur préfèrera, la plupart du temps, obtenir la deuxième car elle aboutit à la mise en évidence de l'effet recherché.

Une autre raison, si l'on suit ce raisonnement, pour laquelle le chercheur préfèrera généralement RH_0 est le fait que, dans ce cas, le risque d'erreur est, non pas nul, mais du moins, connu, maitrisé.

> *Le risque de se tromper en RH_0, (erreur de type I) implique que le modèle H_0 est vrai et donc que les probabilités calculées par ce modèle sont correctes. Je connais donc a priori la probabilité de rejeter H_0 si elle est vraie : nous l'avons définie petite et nommée α.*
>
> (...)

> *Le risque de se tromper en AH_0, (erreur de type II) implique que le modèle H_0 est faux et donc que les probabilités calculées par ce modèle sont incorrectes. Nous nommons cette probabilité β, elle est inconnue et elle peut être grande.*
>
> *β est inconnue car si je fais l'expérience, l'information recherchée est par essence inconnue ; je ne suis même pas certain de son existence ; si elle existe, je ne connais pas l'intensité de son effet sur les données : si l'effet est petit, j'aurai grand risque de ne pas le voir (β serait grande). En conséquence, je ne connais pas a priori la probabilité de me tromper en acceptant H_0 car si elle est fausse, je n'ai pas la valeur des paramètres du modèle qui me permettrait de calculer les probabilités.*
>
> [Depiereux, 2016, p.158]

Il y a donc une importante *asymétrie* entre l'AH_0 et le RH_0, le premier étant fondé sur une *évidence négative* (l'effet n'a pas été mis en évidence) là où le second repose sur une *évidence positive* (l'effet a été mis en évidence).

> *Cette asymétrie entre l'évidence fondée sur l'observation et l'évidence fondée sur l'absence d'observation n'est pas un constat limité aux statistiques, c'est une question de logique qui imprègne l'essence de toute observation scientifique (...). En Biologie, seules les observations basées sur une évidence positive (et en cas de test statistique, basées sur RH_0) sont publiables et considérées comme dignes de foi.*
>
> [Depiereux, 2016, p.158]

> *Le message essentiel est de bien comprendre, à travers les modèles qui décrivent le hasard sur lesquels se basent trois grandes familles de tests statistiques, les comparaisons de variances de moyennes et de fréquences, que l'hypothèse nulle H_0 est (presque) toujours celle que l'expérimentateur veut réfuter et que l'AH_0 constitue dans ce cas un double échec : l'effet escompté n'est pas démontré (évidence négative réfutable) et la probabilité d'erreur (dite de type II et de probabilité β) est inconnue.*
>
> [Depiereux, 2016, p.161]

L'hypothèse qui intéresse le chercheur est donc généralement H_1, mais avant de pouvoir dire que les données corroborent H_1, il convient de prouver que celles-ci étaient très peu probables sous H_0. En effet, la variabilité expérimentale peut très bien générer l'illusion d'un effet réel là où il n'y a que du bruit. Le schéma logique serait alors celui défini à la figure 2.10

Une fois les principes généraux donnés, la logique est appliquée à des données dont le modèle statistique est relativement simple, à savoir, la distribution binomiale.

> *Imaginons une bouteille de culture dans lesquelles se trouvent des cellules qui sont soit vivantes, soit mortes. En fonction du temps de demi-vie, on s'attend à ce qu'il y ait autant de cellules vivantes que de mortes.*

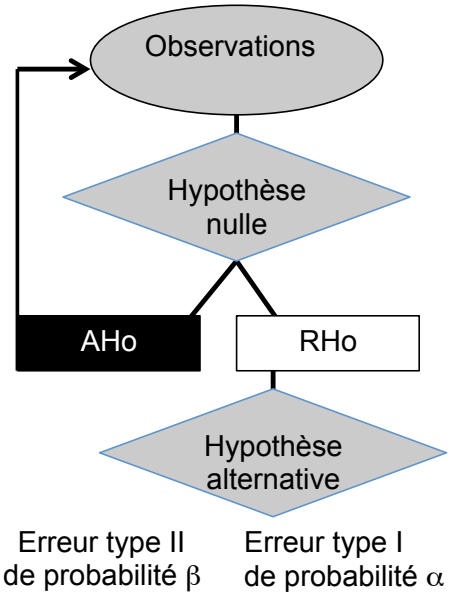


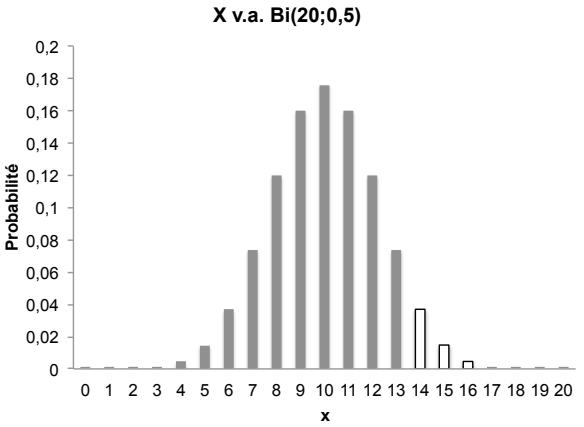
FIGURE 2.10 – Schéma d'un test d'hypothèses [Depiereux, 2016, p.100]

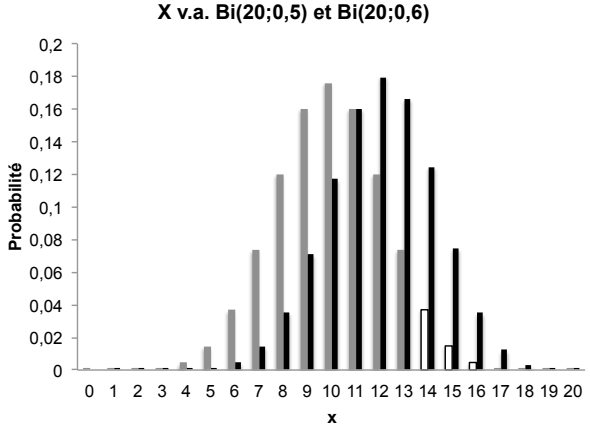
Modèle : si nous observons n cellules prises au hasard et définissons "vivante" comme succès, X le nombre de cellules vivantes est une v.a. Bi(n ; π) dans laquelle π représente la proportion de cellules vivantes et sur laquelle va poser notre hypothèse.

Hypothèse H_0 : pour calculer des probabilités, il faut poser une hypothèse nulle basée sur le temps de demi-vie ; H_0 : $\pi = 0,5$. Notons que théoriquement chaque cellule observée devrait être remise dans la bouteille pour ne pas modifier π.

Hypothèse H_1 : les cellules vivantes étant considérées comme succès, imaginons que l'expérimentateur espère démontrer qu'il y a plus de cellules vivantes et donc que $\pi > 0,5$. Le test est donc unidirectionnel à droite.

Seuil : nous constatons dans les tables de la variable binomiales (annexe) que pour pouvoir définir assez précisément une valeur x correspondant à une probabilité de 5%, n doit être assez élevé. Si $n = 20$, il y a une probabilité de 0,9423 que $X \leq 13$. A partir de 14 cellules vivantes sur 20 cellules observées, H_0 apparaîtra peu probable.





> *Echantillon : procédons à l'observation au microscope de 20 cellules et imaginons que l'on obtienne 12 cellules vivantes.*
>
> *Calcul de la statistique : la statistique observée est simplement X ; nous n'avons pas atteint la limite de 14 et nous acceptons l'hypothèse nulle : AH_0.*
>
> *Conclusion : AH_0 s'accompagne d'une probabilité d'erreur inconnue β, et nous ne sommes guère plus avancés : soit $\pi = 0,5$ et nous avons raison d'accepter, soit $\pi > 0,5$ et nous ne l'avons pas démontré, soit $\pi < 0,5$ et nous n'aurions jamais pu le trouver avec ce test unidirectionnel à droite.*
>
> [Depiereux, 2016, p.161-2]

Cette même logique sera, par la suite, appliquée à d'autres types de données qui font intervenir des modèles statistiques de plus en plus complexes (distribution du χ^2, normale, t de Student, F de Fisher).

L'équation *Variabilité = information + bruit* sera développée pour introduire la comparaison de moyennes basée sur l'analyse des variances (ANOVA), dans laquelle : *Variabilité totale = Variabilité factorielle + variabilité résiduelle.*

Le site d'auto-apprentissage

Tandis que le syllabus donne une présentation détaillée et exhaustive de la matière à connaître, le site d'auto-apprentissage en apporte un résumé agrémenté d'exercices et d'illustrations (statiques ou dynamiques) [Calmant et al., 2017]. Ce site Web est connu pour être largement plébiscité par les étudiants. Dans ce qui suit, nous allons nous intéresser aux pages des modules qui décrivent le principe du test d'hypothèses et les risques d'erreurs (modules 110 et 120, [Calmant et al., 2017]).

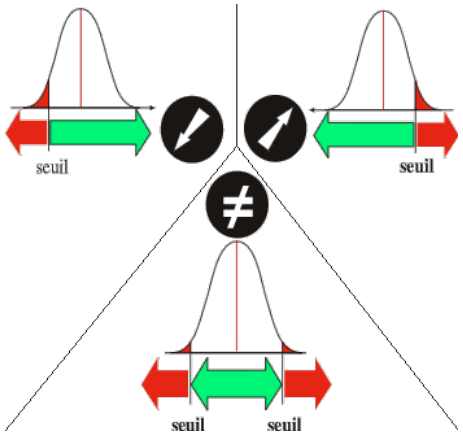
Dans ces pages, le test d'hypothèses est principalement présenté comme un **test de conformité** entre une observation ou une série d'observations et un modèle. Il est d'abord présenté dans un cas de figure relativement simple, la comparaison d'une observation à un standard défini par une loi normale, puis il sera généralisé à d'autres types de tests d'hypothèses.

> *Tout expérimentateur est amené à se poser la question suivante : l'estimation du paramètre d'une population (ex : la moyenne) dans un échantillon est-elle conforme à un modèle établi ? En disant cela, on introduit la notion de test d'hypothèses.*
>
> (...)
>
> *Pour répondre à cette question, l'expérimentateur va devoir définir ARBITRAIREMENT une limite (une frontière) entre la conformité (en vert) et la non conformité (en rouge) à un modèle.*
>
> [Calmant et al., 2017, Module 110]

Trois manières de définir la conformité sont présentées :

1. Le test unidirectionnel à droite : les observations associées aux statistiques les plus importantes sont considérées non conformes ;
2. Le test unidirectionnel à gauche : les observations associées aux statistiques les plus petites sont considérées non conformes ;
3. Le test bidirectionnel : les observations associées aux statistiques les plus extrêmes sont considérées non conformes.

L'expérimentateur peut ainsi définir 3 types de limites arbitraires :



— *test unidirectionnel à droite : Seuil $Z(1-\alpha)$, Confiance $1-\alpha$. Ce cas permet à l'expérimentateur de mettre en évidence une augmentation du paramètre étudié : "L'individu mesuré ou la moyenne de l'échantillon d'individus analysé sont-ils conformes ou plus grands que prévus par le modèle ?"*

— *test unidirectionnel à gauche : Seuil $Z\alpha$, Confiance $1-\alpha$. Ce cas permet à l'expérimentateur de mettre en évidence une diminution du paramètre étudié : "L'individu mesuré ou la moyenne de l'échantillon d'individus analysé sont-ils conformes ou plus petits que prévus par le modèle ?"*

— *test bidirectionnel : Seuil $Z(\alpha/2)$ et $Z(1-\alpha/2)$, Confiance $1-\alpha$. Ce cas permet à l'expérimentateur de mettre en évidence une diminution ou une augmentation (et donc une différence) du paramètre étudié : "L'individu mesuré ou la moyenne de l'échantillon d'individus analysé sont-ils conformes ou non (soit plus grands ou plus petits) au modèle ?"*

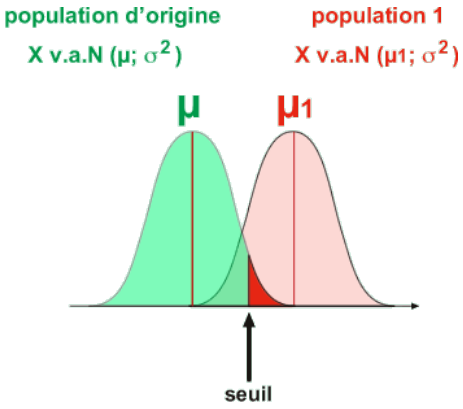
[Calmant et al., 2017, Module 110]

Le modèle auquel on compare l'observation sera donc divisé en **deux zones** : la zone dans laquelle les valeurs sont considérées conformes avec le modèle (en vert sur le graphique) et la zone dans laquelle les valeurs sont considérées comme non-conformes avec le modèle (en rouge sur le graphique).

Module 110 - 3. La zone correspondant à l'erreur de type I (en rouge)

Elle constitue une zone de faible probabilité. En général, elle équivaut à 5%, 1% voire 0,1% de la surface totale sous la courbe de Gauss.

Pour qu'une mesure ou une moyenne se retrouve dans cette zone, il faut que sa valeur soit très éloignée du centre de la distribution μ (ou 0 si on travaille avec des paramètres réduits) au point de dépasser la valeur seuil x_a. Si tel est le cas, il y a beaucoup de chances que cette mesure (ou cette moyenne) n'ait pas été obtenue par hasard. Il est fort probable que la mesure (ou la moyenne) provienne d'une autre population que celle prévue par le modèle H_0 (population 1). En décidant que le modèle H_0 est invalide, il y a cependant alpha % de chances de se tromper. Cet alpha est très faible et donc le risque encouru est mineur.



Module 110 - 4. La zone de confiance (en vert)

Une valeur comprise dans cette zone de confiance est considérée par l'expérimentateur comme une valeur tout à fait conforme au modèle H_0 décrivant la population d'origine centrée sur la moyenne μ.

Cette zone représente 95%, 99% voire 99,9% de la surface de la courbe de Gauss. La distance qui sépare μ d'une valeur observée dans cette zone n'est pas suffisante pour être considérée comme non conforme (car inférieure à la distance séparant μ du seuil de signification).

Dans cette zone, l'expérimentateur doit admettre que la mesure (ou la moyenne) est conforme à la population centrée sur μ. Il n'a pas réussi à démontrer le contraire. Ce n'est pas pour autant que le modèle H_0 est validé mais la valeur obtenue n'est malheureusement pas située en dehors des limites de la zone de confiance, ce qui ne permet pas de postuler que l'échantillon proviendrait d'une autre population, centrée sur M.

(...)

[Calmant et al., 2017, Module 110]

Le modèle auquel on confronte les observations est appelé **modèle** H_0, il repose sur l'hypothèse selon laquelle l'effet étudié n'existe pas, les écarts entre les observations et l'hypothèse H_0 sont uniquement le fruit du hasard.

Si on rejette ce modèle, c'est qu'un autre modèle est plus correct, c'est que l'effet étudié doit exister, être différent de zéro, que les écarts entre les observations et l'hypothèse H_0 ne sont pas uniquement le fruit du hasard.

Dans un test d'hypothèses, il y a donc **deux hypothèses** en jeu : l'hypothèse nulle, H_0, que l'on cherche à rejeter et l'hypothèse alternative H_1, que l'on cherche à mettre en évidence par élimination de l'hypothèse nulle. Cette hypothèse alternative n'est généralement pas connue de manière précise.

Sous chacune de ces hypothèses, les observations ont une certaine probabilité de tomber dans la zone de conformité, ce qui permet de définir les risques d'erreurs :

- sous H_0 :
  - la probabilité d'être dans la zone de conformité (la confiance) est fixée à une valeur élevée, généralement 95% ;
  - la probabilité d'être hors de la zone de conformité (l'erreur α) est fixée à une valeur petite, généralement 5%.
- sous H_1 :
  - la probabilité d'être dans la zone de conformité s'appelle l'erreur β et n'est généralement pas connue vu que le modèle H_1 n'est pas connu précisément ;
  - la probabilité d'être hors de la zone de conformité s'appelle la puissance (1-β) et n'est pas connue non plus.

Il existe donc **deux types de conclusions** possibles à l'issue d'un test d'hypothèses :

Les observations peuvent se situer dans la zone de conformité, ce qui nous amène à accepter H_0 (AH_0), ce qui serait la bonne décision à prendre si H_0 était correct et serait une erreur si H_1 était correct. Vu que le modèle H_1 n'est pas précis, l'erreur β n'est pas connue et donc il est impossible de savoir, si H_1 était vraie, quelle serait la probabilité d'AH_0. Pour cette raison, lorsque l'on AH_0, on ne rejette pas H_1. On dira que les observations sont compatibles avec le modèle H_0 ou que l'effet n'a pas été mis en évidence.

Si les observations se situent hors de la zone de conformité, on dira que l'on rejette H_0 (RH_0), ce qui serait la mauvaise décision à prendre si H_0 était correct (mais ne devrait arriver que rarement si tel était le cas car α aura été fixé à une valeur faible) et serait la bonne décision si H_1 était correct. Dans ce cas, lorsque qu'on RH_0, on dira que l'effet a été mis en évidence (au seuil α).

Le site Web détaille ensuite, de manière très claire, les étapes à suivre lorsque l'on souhaite réaliser un test d'hypothèses :

1. Identifier le type de distribution théorique qui pourra être utilisé en fonction du type de données ;

2. Poser les hypothèses H_0 et H_1 (et donc fixer la manière de définir la conformité) ;
3. Fixer la valeur de l'erreur α tolérée ;
4. Trouver, dans les tables statistiques, les statistiques seuils définissant la zone de conformité ;
5. Calculer, à partir des observations, la statistique observée ;
6. Situer la statistique observée par rapport à la zone de conformité et tirer la conclusion : AH_0 (les observations sont conformes au modèle H_0, l'effet n'a pas été mis en évidence) ou RH_0 (les observations ne sont pas conformes au modèle H_0, l'effet a été mis en évidence).

Module 110 - 6. Exemple

Un écologiste étudie une population de chauves-souris de l'espèce Grand Rhinolophe. D'après la littérature, il sait que l'envergure de ces chiroptères obéit à une distribution normale dont la moyenne est de 375 mm pour une variance de 225 mm^2. Cet écologiste capture un individu dont la taille est de 350 mm. Cet individu est-il considéré conforme ou bien est-il significativement différent de ce que prévoit le modèle ?

Les hypothèses :

— *H_0 : $\mu 1 = \mu$ (l'individu a une taille conforme)*

— *H_1 : $\mu 1$ différent de μ (l'individu est non conforme)*

Attention : "significativement" qualifie une confiance de 95%, par conséquent le alpha = 5%
Ou encore

— *H_0 : M1 = M*

— *H_1 : M1 > M*

— *H_2 : M1 < M*

Convertir la valeur observée en une valeur réduite : X observé = 350 mm ; z observé = (350-375)/15 (où 15 est l'écart-type de la population) ; z observé = -1,66666667.

Trouver le seuil de signification : on sait que alpha vaut 5%. Le test est bidirectionnel (H_1 ou H_2). Le seuil de signification est donc z alpha/2 pour H_2 et z (1-alpha/2) pour H_1 Dans la table, on ne peut trouver que les probabilités P(Z<z) pour z positif. La borne supérieure de l'intervalle de confiance vaut 1,96. En utilisant la propriété de symétrie la borne inférieure vaut -1,96.

Conclusion :

-1,96<-1,66666667<1,96 c'est-à-dire z observé compris [dans la zone] de confiance. On accepte H_0 (AH_0). Cela veut dire qu'on n'a pas réussi à montrer que l'individu capturé était significativement (alpha de 5%) différent de la normale.

(...)

Une fois le principe et la procédure expliqués sur un cas simple (comparaison d'un individu à un modèle normal), les procédures correspondant aux cas de figure plus complexes sont amenées :

1. Comparaison d'une moyenne d'un échantillon à un modèle normal (de moyenne et écart-type connus) :
2. Comparaison d'une moyenne d'un échantillon à un modèle normal (de moyenne connue mais d'écart-type inconnu) :
3. Comparaison d'une différence de deux moyennes à un modèle normal (échantillons pairés) ;
4. Comparaison d'une différence de deux moyennes à un modèle normal (échantillons indépendants)
5. Comparaison de plusieurs moyennes via l'analyse de la variance ;
6. Comparaison de variances à l'aide de la distribution F de Fisher ;
7. Comparaison de fréquences observés avec les fréquences prédites par un modèle (distribution du χ^2).

2.4.2 Définition d'une praxéologie (D)

A partir de la manière dont le test d'hypothèses est présenté dans le syllabus et sur le site d'auto-apprentissage, nous pouvons définir une quatrième praxéologie (D).

Pour cette praxéologie, le **type de tâche** serait de "Mettre en évidence un effet en présence de bruit".

Pour accomplir ce type de tâche, la **technique** suivante est proposée :

1. Définir H_0 = Absence d'effet et H_1 = Présence d'effet (hypothèse d'intérêt)
   H_1 pourra être de type uni-directionnel ou bi-directionnel selon les contextes mais n'est pas défini précisément, ce qui implique que le risque d'erreur β ne sera pas défini ;
2. Fixer le risque d'erreur α toléré, généralement à 5 % ;
3. Choisir une statistique de test, $d(X)$;
4. Modéliser la distribution de la statistique de test sous H_0, $f(d(X)|H_0)$;
5. Trouver la statistique seuil, $d(x_s)$. Cette statistique sépare les résultats selon qu'ils conduiront à mettre en évidence un effet (zone de rejet de H_0, RH_0) ou non (zone d'acceptation de H_0, AH_0) ;
6. Calculer la statistique observée à partir des données, $d(x_o)$;
7. Utiliser cette statistique pour tirer la conclusion générale :

— soit en situant $d(x_o)$ par rapport aux zones de RH_0 ou de AH_0,
— soit en calculant la P-valeur unilatérale ou bi-latérale associée à cette statistique observée et en la comparant à α [13] ;

Les deux approches conduisent au mêmes conclusions :

— en cas de AH_0 (ou de P-valeur $> \alpha$), on considère que l'effet n'est pas mis en évidence, que les données sont compatibles avec H_0 jusqu'à preuve du contraire. La probabilité de se tromper en AH_0 serait élevée dans ce cas puisque le risque d'erreur β reste inconnu. Il s'agirait donc d'un résultat relativement incertain, peu définitif,
— en cas de RH_0 (ou de P-valeur $\leq \alpha$), on considère que l'effet étudié est mis en évidence, a été démontré. Dans ce cas, la probabilité de se tromper en RH_0 serait petite ($\leq \alpha$) ce qui implique que la conclusion est considérée plutôt fiable.

Comment cette technique est-elle justifiée ? Quelle est la **technologie** sous-jacente ?

Il semble qu'elle n'apparaisse pas de manière explicite, peut-être parce qu'il s'agit d'un cours de statistique appliquée. Sans doute qu'implicitement, la démarche est justifiée avec des éléments du type :

— la justification peut être trouvée, si on la cherche, au niveau du savoir savant ;
— il ne semble pas que l'enseignement soit très différent dans les autres établissements universitaires ;
— cela correspond à ce que l'on voit dans la littérature scientifique.

La technique n'est donc pas démontrée à proprement parler dans le cadre de ce cours. Par contre, on peut noter la présence de simulations permettant de vérifier que le comportement des modèles est bien celui présenté dans le cours. Les simulations sont notamment utilisées pour montrer la puissance qui serait obtenue en fonction de l'importance de l'effet réel à mettre en évidence et en fonction de la variabilité individuelle.

Sur quels **éléments théoriques** cette technique et cette technologie peuvent-elles se fonder ? A nouveau, ceux-ci n'apparaissent pas tels quels dans la présentation de la démarche, ni dans le syllabus ni dans le site d'auto apprentissage, mais on pourrait *a posteriori* avancer les deux éléments suivants.

D'une part, on trouve derrière cette démarche l'idée selon laquelle il est parfois difficile de distinguer un effet expérimental du simple bruit. Dès le moment où il existe une variabilité individuelle, il y a un risque de résultat faussement positif (mettre en évidence un effet alors que celui-ci n'existe pas) ou faussement négatif (ne pas mettre en évidence un effet qui, en réalité, existe). Le test d'hypothèses est vu comme un instrument permettant de ne pas affirmer trop souvent qu'un effet existe, un moyen de contrôler le risque de faux positifs.

D'autre, part, on peut noter que la justification de cette démarche passe par une référence

13. Notons, cependant, que la P-valeur n'est que brièvement mentionnée dans le syllabus. Cette manière de faire serait donc cohérente avec la praxéologie D mais n'est que peu appliquée.

TABLE 2.7 – **Résumé de l'analyse du savoir enseigné en termes de praxéologie**

Niveau praxéologique	**Praxéologie** D
Type de tâche	Mettre en évidence un effet en présence de bruit
Technique	1. Poser H_0 = Absence d'effet et H_1 = Présence d'effet 2. Fixer α 3. Choisir $d(X)$ 4. Modéliser $f(d(X)\|H_0)$ 5. Calculer $d(x_s)$, délimitant les zones d'AH_0 et de RH_0 6. Calculer $d(x_o)$ 7. Calculer la P-valeur associée à $d(x_o)$, si $\leq \alpha$: effet mis en évidence, faible probabilité d'erreur, si $> \alpha$: effet non mis en évidence, probabilité d'erreur inconnue.
Technologie	Existe en d'autres institutions Utilisation de simulations
Théorie	Contrôle du bruit Asymétrie entre réfutation définitive et acceptation provisoire

indirecte à l'idée de corroboration et de réfutation qui définit la démarche scientifique selon Popper. En effet, on trouve dans le raisonnement enseigné l'idée que l'on n'accepte pas l'hypothèse nulle (de la même manière que l'on ne prouve jamais définitivement une hypothèse en sciences). Cela amène à des conclusions prudentes lorsque les résultats ne contredisent pas H_0 : on dira que l'effet *n'est pas mis en évidence* (ce qui ne signifie pas qu'H_0 soit vraie).

A l'inverse, les résultats peuvent réfuter définitivement une hypothèse. Cela se traduit, dans cette version du test d'hypothèses, par l'idée que le rejet de l'hypothèse nulle possède un statut différent de l'acceptation de l'hypothèse nulle, qu'il y a une asymétrie entre AH_0 et RH_0. Cette asymétrie entre corroboration provisoire et réfutation définitive se retrouve derrière les concepts d'évidences positive (à caractère définitif) et négative (à caractère provisoire).

Le test d'hypothèses est donc vu comme un instrument qui reproduit certaines caractéristiques de la démarche scientifique.

2.4.3 Discussion

Comparaison avec les praxéologies A, B et C

Ayant ainsi décomposé le savoir enseigné en une praxéologie (voir tableau 2.7, p.84), nous pouvons aisément la comparer avec les praxéologies identifiées dans le savoir savant, à savoir les praxéologies A, B et C_1 (voir tableau 2.4, p. 59).

On note ainsi que la praxéologie D diffère des praxéologies A, B et C_1, tout en empruntant des éléments à chacune d'elles.

En effet, **par rapport à (la praxéologie)** A, on note que D partage

1. L'idée de corroboration/réfutation d'une hypothèse ;
2. Le concept de P-valeur ;
3. Le recours à $f(d(X)|H)$;
4. Le rejet de H en présence de petites P-valeurs.

Mais on peut noter les différences suivantes :

D'une part, l'hypothèse initiale est H_0, hypothèse selon laquelle l'effet étudié n'existe pas, qui est souvent l'inverse de l'hypothèse d'intérêt. Au lieu d'être dans une démarche de corroboration progressive de H (comme dans A), on se retrouve dans une démarche de démonstration par l'absurde puisque l'on réfute H_0 pour prouver H_1. Ce léger changement au niveau de l'hypothèse à la base du test modifie la philosophie de celui-ci.

D'autre part, le concept de P-valeur n'est pas exactement le même dans D où les P-valeurs peuvent être calculées de manière uni-directionnelle ou bi-directionnelle selon l'hypothèse d'intérêt (H_1). Dans A, une petite P-valeur indique des résultats qui s'écartent du modèle construit sous H, tandis que dans D, une petite P-valeur indique que les résultats s'écartent du modèle construit sous H_0 *en direction de* H_1. Dans A, la P-valeur est donc un instrument de mesure du niveau auquel les observations corroborent H tandis que, dans D, elle devient un instrument indiquant quant il faut préférer H_1 à H_0.

Par rapport à B, la praxéologie D emprunte les éléments suivants :

1. L'idée de faire un choix, objectif, entre deux conclusions ;
2. Le concept de risque d'erreurs α et β ;
3. Le concept de zone d'acceptation d'une hypothèse ;
4. La possibilité de construire des tests uni-directionnels.

Cependant, on note les différences suivantes :

La praxéologie D ne s'inscrit pas entièrement dans une méthodologie de choix entre deux hypothèses concurrentes et se justifie plutôt par référence indirecte à l'idée de corroboration/réfutation d'une hypothèse.

Le risque d'erreur β dans la praxéologie D n'est généralement pas défini puisque l'ampleur de la différence attendue entre H_0 et H_1 n'est pas précisée.

Dans B, on trouve une zone d'acceptation de H_0 et une zone d'acceptation de H_1 tandis que dans D, on trouve une zone d'acceptation de H_0 et une zone de rejet de H_0. Cela s'explique, une fois encore, par la recherche d'une preuve par la réfutation définitive.

Enfin, bien que D ne s'inscrive pas dans un cadre probabiliste bayésien, on peut noter **un point commun avec les praxéologies** C, et donc C_1 notamment : c'est le fait d'affirmer que l'on connait la probabilité de se tromper à l'issue du test. Dans D, en effet, lorsque l'on rejette l'hypothèse nulle, on affirme que la conclusion est fiable car la probabilité de se tromper en rejetant l'hypothèse nulle est faible. Ce faisant, on assimile la P-valeur à une probabilité *a posteriori*. En effet, alors que la P-valeur $= P(d(X) > dx_0|H_0)$ [14] , c'est-à-dire la probabilité d'observer certains résultats sous H_0, la probabilité *a posteriori* se définirait comme $Posterior = P(H_0|d(x_o))$, c'est-à-dire la probabilité que H_0 soit vraie ayant observé un certain résultat. Alors qu'il parait justifié de parler de probabilité d'erreur à l'issue de la conclusion d'un test statistique bayésien, cela ne l'est pas à l'issue d'un test d'hypothèses.

On le voit, D est une praxéologie différente de celles que nous avons décrites au niveau du savoir savant. En cela, on pourrait dire qu'il est incorrect d'affirmer que l'on enseigne B (la praxéologie initialement associée au test d'hypothèses) lorsque l'on enseigne D (la praxéologie enseignée, appelée également test d'hypothèses).

Mais surtout, cette observation amène de nombreuses questions :

Comment les praxéologies A, B et C du savoir savant se sont-elles transformées en une praxéologie D dans le savoir enseigné ?

Aurait-on pu enseigner A ? ou B ? ou C ? Dans quelle mesure est-il possible d'importer ces praxéologies dans le savoir enseigné sans les altérer ?

Ou, plus généralement : quelles sont les contraintes institutionnelles qui s'appliquent au savoir enseigné et qui peuvent expliquer l'importante transposition didactique que nous venons de décrire ?

14. Dans le cas de tests bi-directionnels, et avec $d(X)$ mesurant l'écart à H_0, comme c'est le cas avec la statistique du χ^2 par exemple

Recherche des contraintes institutionnelles

Identifier des contraintes institutionnelles pesant sur le savoir enseigné est loin d'être une tâche aisée et nous ne pourrons certainement pas prétendre les avoir identifiées toutes, d'autant que ces contraintes varient dans l'espace et dans le temps. Ci-dessous, à partir des principales contraintes que nous avons relevées, nous détaillons les praxéologies qu'elles nous semblent favoriser.

Selon Verret (1975, cité dans [Chevallard, 1991]) :

> *"Une transmission scolaire bureaucratique suppose quant au savoir*
>
> 1°– *la division de la pratique théorique en champs de savoir délimités donnant lieu à des pratiques d'apprentissage spécialisées – c'est-à-dire la désyncrétisation du savoir.*
>
> 2°– *en chacune de ces pratiques, la séparation du savoir et de la personne – c'est-à-dire la dépersonnalisation du savoir.*
>
> 3°– *la programmation des apprentissages et des contrôles suivant des séquences raisonnées permettant une acquisition progressive des expertises – c'est-à-dire la programmabilité de l'acquisition du savoir.*
>
> *Elle suppose quant à la transmission :*
>
> 1°– *la définition explicite, en compréhension et en extension, du savoir à transmettre – c'est-à-dire la publicité du savoir.*
>
> 2°– *le contrôle réglé des apprentissages suivant des procédures de vérification autorisant la certification des expertises – c'est-à-dire le contrôle social des apprentissages".*

Pour les savoirs mathématiques, comme pour d'autres types de savoirs, ces contraintes générales – désyncrétisation, dépersonnalisation, programmabilité, publicité et contrôle social des apprentissages – agissent au moment de la *mise en texte du savoir* [Chevallard, 1991].

Suivant ces contraintes générales, nous pouvons déjà constater que les praxéologies A, B et C telles que nous les avons définies seraient assez mal adaptées à l'enseignement dans la mesure où elles mettent en œuvre des éléments de jugement subjectif qui sont, par nature, difficilement programmables, qu'il est difficile d'expliciter, de rendre public et d'évaluer.

En effet, l'interprétation d'une P-valeur selon Fisher fait intervenir le jugement du scientifique qui doit intégrer *a posteriori* les éléments de contexte afin de tirer une conclusion générale. Mais comment enseigner la manière dont il convient d'interpréter un résultat en fonction du contexte ? Comment programmer l'acquisition de tels savoirs et, surtout, comment les évaluer ?

De la même manière, la définition des hypothèses dans le test d'hypothèses selon Neyman et Pearson implique une connaissance du contexte pour choisir l'hypothèse de référence et celle

qui constitue l'alternative, pour définir ce que seraient des risques acceptables dans le contexte. Cette démarche *a priori* requiert également un certain discernement.

La praxéologie C_1, quant à elle, fait également intervenir de la subjectivité au moment de la définition du niveau de croyance *a priori* dans les hypothèses en jeu.

Au contraire, la praxéologie D a le "mérite" de ne pas engager le discernement, la subjectivité. A partir d'une série de données, tout chercheur, tout étudiant, doit arriver à définir la même hypothèse nulle, et doit parvenir à la même conclusion quant à la mise en évidence ou non d'un effet expérimental. Cela est particulièrement vrai lorsque les jeux de données sont simplifiés et stéréotypés [15]. Ce type de démarche peut, dès lors, être explicité, son apprentissage peut être programmé, l'acquisition des connaissances peut être évaluée, bref cette praxéologie peut être enseignée.

Elle peut ainsi prendre la forme d'un algorithme, de "recettes de cuisine", dont la maitrise est assez facile à évaluer mais elle peut poser problème à ceux qui considèrent que l'analyse de données dans le but de construire des connaissance scientifiques est une entreprise nécessairement empreinte de subjectivité.

En somme, l'application des contraintes générales énoncées par Verret à notre contexte nous amène à penser qu'une praxéologie aura d'autant plus de facilités à survivre dans une institution d'enseignement qu'elle sera **objective**, qu'elle n'impliquera pas la mise en œuvre d'un discernement.

Au-delà des contraintes qui agissent, de manière générale, sur la *mise en texte* d'un savoir mathématique, on notera également qu'il existe certaines contraintes spécifiques à l'enseignement de la biostatistique aux étudiants des filières biomédicales à l'université de Namur :

- cet enseignement a une visée appliquée : l'étudiant est censé développer une vision critique des méthodes statistiques utilisées dans la littérature biomédicale ;
- les objectifs sont ambitieux par rapport au volume horaire disponible (4 ECTS [16]) ;
- le public est composé de cohortes importantes d'étudiants ;
- actuellement et au sein de l'université de Namur, les enseignants ont une formation initiale scientifique et non mathématique.

Alors que, sur base de l'analyse du savoir savant, nous aurions pu décrire le savoir enseigné idéal, hors contraintes, comme étant composé de A, B, C à utiliser selon les contextes, de manière nuancée, complémentaire et critique, nous observons que les contraintes spécifiques liées au savoir enseigné vont favoriser l'émergence d'une praxéologie **unique** et **basée sur les pratiques de référence** plutôt que sur le savoir savant.

15. Comme nous l'avons fait avec l'exemple générique utilisé pour comparer les praxéologies A, B et C, p62.
16. Censé correspondre à une formation de 120 h comprenant cours théoriques, travaux pratiques, étude et évaluation

La réduction à une seule praxéologie enseignée nous semble, en effet, difficilement évitable au vu du volume horaire dévolu à cet enseignement. Pourtant, la présentation d'une seule méthode pour tester des hypothèses statistiques est, en soi, problématique. Comme dit le proverbe, "*If all you have is a hammer, everything looks like a nail*".

Par ailleurs, le profil de l'enseignant et des assistants − scientifiques plutôt que mathématiciens − ainsi que la visée de cet enseignement − préparer l'étudiant à comprendre la pratique actuelle d'analyse de données en recherche médicale − favorisent vraisemblablement les praxéologies, d'une part, dans lesquelles la technologie repose sur des illustrations et des simulations plutôt que sur des démonstrations − technologies qui ont leurs limites [Dagnelie, 2010]− et, d'autre part, pour lesquelles la proximité avec les pratiques *effectives* en recherche scientifique est visiblement plus importante que la proximité avec les praxéologies qui peuvent exister au sein d'un savoir savant.

Ainsi, nous pensons que les contraintes qui s'appliquent actuellement au savoir enseigné sont de nature à favoriser l'enseignement d'une praxéologie **unique**, **objective** et **proche de la pratique**.

En particulier, le besoin, dans un cours de statistique appliquée, d'enseigner "ce qui se pratique" est, en effet, une contrainte majeure dans la détermination d'une praxéologie enseignée. Dans ce contexte, mieux comprendre la transposition didactique à l'oeuvre passe donc nécessairement par un examen approfondi de ces pratiques, par l'identification des praxéologies qui peuvent y exister et par l'analyse des contraintes institutionnelles qui les déterminent.

2.5 Pratiques sociales de référence

En décrivant la transposition didactique de savoirs principalement mathématiques, Chevallard se focalise sur l'étude des altérations qui surviennent lorsque le savoir savant est transposé dans une forme scolaire.

Ayant travaillé sur des disciplines techniques dans l'enseignement secondaire, Martinand propose d'étudier les écarts entre les pratiques scolaires et celles qui leur servent de références dans le milieu professionnel ou extra-professionnel, les *pratiques sociales de référence* [Martinand, 1989]. En effet, les savoir savant n'est pas toujours celui qui sert implicitement de référence aux savoirs enseignées.

L'idée n'est pas de vérifier que l'on observe bien une stricte identité entre les pratiques scolaires et les pratiques sociales de référence, mais plutôt, à l'image de ce que propose Chevallard, d'interroger la nécessaire transposition que les savoirs subissent nécessairement lorsqu'ils s'adaptent à une institution différente. Le concept de pratiques sociales de référence vise à poser la question de la référence des savoirs scolaires – Ne font-ils pas uniquement référence à

eux-mêmes ? A quelles pratiques fait-on référence dans notre enseignement ? – et à questionner la pertinence des écarts entre les uns et les autres.

Dans le cas des pratiques d'inférence statistiques enseignées aux étudiants des filières biomédicales à l'Université de Namur, les pratiques sociales de référence sont explicitement constituées des pratiques d'inférence statistique retrouvée dans le milieu de la recherche biomédicale et donc dans la littérature biomédicale.

Comprendre la transposition didactique des outils d'inférence statistique nécessite donc de prendre en compte ces pratiques sociales de référence et d'analyser en quoi elles se distinguent du savoir savant, d'une part, et du savoir enseigné, d'autre part.

Dans cette section, nous chercherons, tout d'abord, à définir les notions d'inférence statistique les plus courantes dans la littérature biomédicale.

Nous verrons que le rapport des chercheurs à ces objets pose problème et est, en partie, à l'origine de la crise de reproductibilité des résultats scientifiques.

Nous tenterons, ensuite, de définir une praxéologie associée à la pratique de l'inférence statistique par les scientifiques. Pour ce faire, nous nous appuierons sur les écrits d'auteurs qui dénoncent certaines utilisations courantes mais abusives des tests statistiques en recherche.

Enfin, nous tenterons d'identifier les facteurs qui ont contribué à l'émergence et à la diffusion de ce type de pratiques.

2.5.1 Identification des savoirs utilisés

La première question à laquelle nous allons tenter d'apporter une réponse est : "Quelles notions liées aux tests statistiques rencontre-t-on dans la littérature biomédicale ?"

Pour le savoir, nous réaliserons un tirage aléatoire d'articles scientifiques et nous y analyserons les notions d'inférence statistique rencontrées.

Sélection des articles

Nous avons d'abord choisi, arbitrairement, trois revues prestigieuses dans la littérature biomédicale : *Lancet*, *Journal of the American Medical Association* (JAMA) et *New England Journal of Medicine* (NEJM). Ces revues font partie des revues les plus influentes dans la littérature biomédicale [17] ce qui peut raisonnablement nous faire penser que les méthodes qui y sont utilisées sont assez bien acceptées dans le reste de la littérature biomédicale. Par ailleurs, le fait que ces revues effectuent une sélection rigoureuse des articles pouvant y être publiés diminue

17. A titre d'information les facteurs d'impacts 2018 sont, respectivement de 59, 51 et 71

TABLE 2.8 – **Identifiants (PubMed ID) des articles utilisés dans cette enquête**.

NEJM			JAMA			Lancet		
12944569	12556542	10816184	12771113	19755697	11368699	12133656	17617271	12814710
11309635	10675425	12601075	14679271	20805623	12759324	15964448	18926569	12433513
20860505	20375405	16525139	12095381	19351943	20009055	19932356	17630037	16829296
10770982	14695409	16481635	15769967	19797474	10770145	17980734	16503464	10972371
19336502	16540614	16291982	15928285	20736470	15173148	16338450	12814712	17905167
10874061	15306665		11427138	18165667		20417856	12126818	
17050889	11106715		10697060	12243636		12547542	20580423	
16306520	11759643		19724041	16757723		11145488	20801495	
16971716	20573923		19017911	14679270		12648967	10791374	
19671655	18768944		14612479	18378631		18502299	19570573	

le risque de tomber sur un article dans lequel les notions statistiques sont mal utilisées.

Ensuite, parmi ces revues, nous avons choisi de nous focaliser sur les articles qui décrivent des résultats d'essais cliniques randomisés (*Randomized clinical trials*). Parmi les différents types d'investigations en recherche biomédicale, les essais randomisés sont généralement considérés comme les expériences les plus fiables, celles qui apportent le niveau de preuve le plus élevé. Seules les méta-analyses, recherches systématiques dans lesquelles les résultats de plusieurs essais cliniques sont agrégés, peuvent être considérées plus fiables que les essais cliniques randomisés individuels mais nous ne les prendrons pas en considération ici car ces études mettent en œuvre des méthodes statistiques bien particulières qui ne reflètent pas les outils généralement utilisés par les chercheurs dans le domaine biomédical.

Enfin, nous avons restreint la période d'analyse à l'intervalle allant du 1er janvier 2000 au 31 décembre 2010, ce qui correspond aux pratiques avant le début de cette thèse.

Ces critères de sélection nous ont donné une liste de publications contenant 649 articles du *JAMA*, 1004 pour le *Lancet* et 1063 articles pour le *NEJM*. Dans chacune de ces trois listes, nous avons sélectionné aléatoirement 25 articles. Parmi ces 75 articles, 7 ont été exclus car il s'agissait de commentaires à propos d'essais cliniques et non pas d'essais cliniques à proprement parler, 1 a été exclu car l'article a été rétracté depuis sa publication et 1 a été exclu car il s'agissait d'un article bibliométrique et non médical. Ces 9 articles exclus ont été remplacés de manière à obtenir trois séries de 25 articles (voir tableau 2.8).

Résultats

Dans chacun de ces articles, nous avons déterminé la présence des notions suivantes :
— la P-valeur ;
— l'intervalle de confiance (IC) ;
— la puissance ;
— les méthodes d'inférence bayésienne (*Bayes factor* ou probabilité *a posteriori*).

Nous avons fait la distinction entre les notions qui se trouvent dans l'article mais pas dans le résumé, et celles qui se trouvent à la fois dans l'article et dans le résumé. Nous avons, en effet, considéré que le fait qu'une information soit reprise dans le résumé témoigne de l'importance qu'elle a au yeux des chercheurs. Les résultats sont présentés à la figure 2.11.

On observe que la P-valeur est présente dans 99 % [intervalle de confiance binomial à 95 % : 93 à 100 %] des articles et dans 84 % des résumés [74 % à 91 %]. On trouve un intervalle de confiance dans 88 % des articles [78 % à 94 %] et 64 % des résumés [52 % à 75 %]. La notion de puissance se rencontre quant à elle dans 81 % des articles [71 % à 89 %] mais seulement dans 1 % des résumés [0% à 7 %]. Enfin, les méthodes d'inférence bayésienne se trouvent dans 1 % des articles [0 % à 7 %] et dans 0 % des résumés [0 % à 5 %].

Discussion

Différents éléments peuvent être retirés de cette analyse de la littérature.

Actuellement, ce sont presque exclusivement des méthodes d'analyse inférentielle non-bayésienne qui sont utilisées. Il se peut que dans certains domaines spécifiques (essais cliniques dits de phase I par exemple), l'inférence bayésienne occupe une place plus importante mais globalement, on peut considérer qu'elle n'est utilisée que de manière anecdotique.

On voit que, parmi les outils d'inférence fréquentistes, la P-valeur occupe une place prépondérante, incontournable. Cette notion se rencontre dans plus de quatre résumés sur cinq et dans quasiment tous les articles. On peut noter que cette P-valeur est généralement donnée de manière exacte et plutôt que sous la forme ($P < 0.05$) et correspond généralement à une P-valeur bidirectionnelle.

La notion d'intervalle de confiance vient très souvent compléter la P-valeur, quelquefois elle la remplace dans le résumé. Dans environ un essai clinique sur cinq, cette notion est absente. C'est notamment le cas lorsque l'analyse des données repose sur des méthodes non-paramétriques.

Si le concept de P-valeur propre au raisonnement de Fisher est prépondérant, il ne semble pas incompatible avec le fait d'utiliser la logique du test d'hypothèses dans l'élaboration du

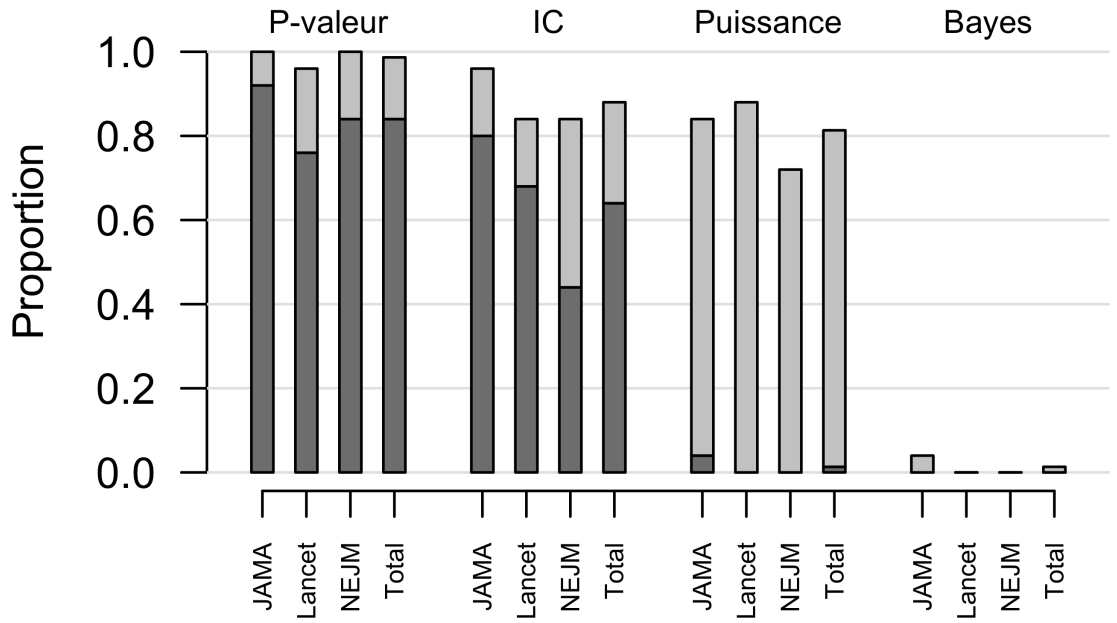


FIGURE 2.11 – **Proportion d'articles dans lesquels apparaissent les notions d'intérêt**. La distinction a été faite selon que la notion apparaissait dans le résumé (gris foncé) ou dans le reste de l'article (gris clair). La proportion indiquée a été calculée sur 25 articles par journal.

design de l'expérience. Cependant, à quelques exceptions près, on ne retrouve pas la logique du test d'hypothèses en dehors des considérations liées à la taille d'échantillon.

2.5.2 Un rapport au savoir qui pose problème

Si on compare ce que nous avons observé dans la littérature biomédicale à ce qui a été décrit au niveau du savoir savant, on constate, à première vue, que les concepts statistiques sont similaires. On notera simplement que les concepts bayésiens restent actuellement peu utilisés.

Cependant si on y regarde d'un peu plus près, au-delà de la simple présence ou absence d'un concept statistique, on constatera que les rapports à ces concepts que l'on rencontre dans les pratiques de référence diffèrent de ceux décrits dans le savoir savant. Ou, pour être plus précis, il semble exister, au sein de ces pratiques, une praxéologie qui diffère substantiellement de celles que l'on retrouve dans le savoir savant, et qui, si elle n'est pas la seule praxéologie rencontrée dans la littérature biomédicale, est suffisamment répandue et problématique pour être considérée comme une des causes de la crise de reproductibilité des résultats scientifiques.

Le rapport des scientifiques aux outils d'inférence statistique pose question depuis longtemps. En 2001 déjà, Thompson recensait pas moins de 402 articles dénonçant l'abus d'utilisation des outils d'inférence fréquentistes dans la recherche scientifique, et en particulier dans les études observationnelles.

Depuis, le rapport des scientifiques à ce savoir ne semble pas avoir changé. Au contraire, les données devenant de plus en plus accessibles, de plus en plus nombreuses, le besoin de maîtrise des outils d'inférence statistique s'est considérablement accru alors que les capacités des chercheurs à analyser les données ne se sont pas forcément améliorées [Peng, 2015].

> *One result of this is an epidemic of poor data analysis, which is contributing to a crisis of replicability and reproducibility of scientic results* [Peng, 2015].

Plutôt que de s'être résorbés, les problèmes liés à la maîtrise des outils d'inférence statistique se sont donc amplifiés et cela a contribué à la crise de la reproductibilité des résultats scientifiques.

2.5.3 Crise de reproductibilité

Prise de conscience de l'existence d'un problème

Dans le courant des années 2000, deux éléments viennent ébranler la confiance des scientifiques dans le caractère reproductible des études publiées dans la littérature [Peng, 2015].

D'une part, dans de nombreux domaines, des tentatives de reproduction d'expériences importantes dans leur domaine se sont soldées par des échecs [Peng, 2015] : en génomique appliquée au cancer [Potti et al., 2011] en bioinformatique [Baggerly and Coombes, 2009] ou encore en économie [Herndon et al., 2014].

D'autre part, dans son très influent article intitulé " *Why most published research findings are false*", Ioannidis (2005) apporte des éléments théoriques soutenant que la plupart des résultats publiés sont faux [Ioannidis, 2005b].

Son raisonnement est le suivant : il cherche à connaitre la probabilité qu'un résultat de recherche soit vrai. Pour cela, il part d'une approche bayésienne (voir chapitre 1) dans laquelle il cherche à déterminer la probabilité que l'hypothèse alternative soit vraie *a posteriori*. Il étudie, par simulations, l'influence de deux variables sur cette probabilité *a posteriori* : la puissance de l'étude $(1-\beta)$ et le biais. Le biais peut se définir comme l'écart moyen entre la vraie valeur (la valeur du paramètre au niveau de la population théorique) et la valeur mesurée au cours d'une expérience. Pour le mesurer précisément, il faudrait donc d'une part, connaître la vraie valeur et d'autre part, disposer d'un grand nombre de répétitions d'une certaine expérience visant à estimer cette valeur. Il serait alors égal à la différence entre la vraie valeur et la moyenne des

estimations. De nombreux facteurs peuvent provoquer un biais dans le processus d'estimation, c'est notamment le cas du design expérimental, de la manière d'analyser les données ou encore la manière dont on distingue les résultats que l'on diffuse ou non.

De ces simulations, il déduit que la majorité des résultats publiés dans la littérature scientifiques sont probablement faux et que la probabilité qu'un résultat soit faux est d'autant plus grande que :

— peu d'études ont été publiées sur le sujet ;
— l'effet observé est léger ;
— le nombre de relations étudié était important ;
— la liberté dans l'analyse des données était grande ;
— les conflits d'intérêts sont potentiellement importants ;
— le sujet est "brûlant".

Le dernier point peut sembler entrer en contradiction avec le premier. Ioannidis explique que lorsqu'un sujet est brûlant [18], beaucoup d'équipes travaillent en même temps sur le même sujet et que la probabilité *a posteriori* de chaque résultat individuel est alors réduite. En effet, dans ces conditions, le biais augmente car chaque équipe tend à se focaliser sur la diffusion des résultats qui semblent les plus impressionnants. Cette phase peut alors être suivie d'une réfutation forte des premiers résultats publiés, phénomène que l'auteur a nommé le *Proteus phenomenon* [Ioannidis, 2005b].

Selon lui, pour réduire la proportion de résultats faussement positifs dans la littérature, il conviendrait de (1) favoriser les études de grande taille (en particulier dans les domaines ou la probabilité *a priori* qu'un effet existe est importante et avec des designs qui permettent de fermer un champ de recherche en cas d'étude négative) ; (2) ne pas focaliser l'attention sur des études individuelles mais plutôt sur des synthèses de plusieurs études, des méta-analyses ; (3) favoriser l'enregistrement des essais cliniques dans des bases de données générales permettant ainsi de se faire une idée de l'éventuel biais de publication [19] ; et (4) utiliser les outils d'inférence bayésienne dans la conduite de la recherche.

Par ailleurs, dans un autre article, Ioannidis apporte des preuves empiriques à son hypothèse [Ioannidis, 2005a]. Dans un premier temps, il a sélectionné 49 essais cliniques qui ont été largement cités dans la littérature (au moins 1000 fois chacun) et publiés dans une des trois grandes revues médicales (*New England Journal of Medicine*, *Journal of the American Medical*

18. Pensons à l'épidémie de Covid-19 qui nous préoccupe au moment d'écrire ces lignes

19. Biais qui survient lorsque toutes les études faites dans un domaine ne sont pas publiées et que la probabilité d'être publiée est liée au résultat obtenu. En d'autre termes, si les études montrant un effet positif et significatif d'un traitement ont plus de chances de se voir publiées et diffusées, alors l'effet du traitement dans la littérature sera une sur-estimation de l'effet réel.

Association et *The Lancet*) ou dans une revue spécialisée à haut facteur d'impact. Il a ensuite comparé les résultats de ces 49 études aux études publiées ensuite ayant un design expérimental au moins aussi bon que l'étude publiée initialement.

Dans ces 49 études extrêmement influentes dans leur domaine, 45 affirmaient que l'intervention étudiée était efficace. De ces 45 études, 7 (16 %) ont été contredites, 7 (16 %) présentaient un effet qui s'est avéré moins important dans les études postérieures, 20 (44 %) ont été répliquées et 11 (24 %) n'ont pas pu être comparées à d'autres études [Ioannidis, 2005a].

Il montre donc que, même en s'intéressant aux études parmi les plus influentes au niveau médical, publiées dans les plus grandes revues, on constate que seules 44 % ont pu être répliquées quelques années plus tard. Environ un quart des études n'ont pas été "challengées" et un tiers des études donnaient une estimation fausse (allant dans le mauvais sens) ou exagérée.

Place des pratiques d'inférence statistique dans la crise de reproductibilité

La prise de conscience qu'une part importante des résultats d'études scientifiques ne peuvent être reproduits a conduit à ce qui a été appelé la *reproducibility crisis* ou la crise de la reproductibilité des résultats scientifiques. Dans une enquête conduite par la revue *Nature*, 1576 scientifiques se sont prononcés concernant cette crise de reproductibilité. Plus de 70 % d'entre eux affirment avoir échoué à reproduire l'expérience de quelqu'un d'autre et plus de la moitié a déjà échoué à reproduire une de ses propres expériences [Baker, 2016]. Nonante pour cent d'entre eux admettent qu'il existe une crise de reproductibilité en sciences.

Dans cette même enquête, les chercheurs se sont prononcés sur les facteurs qui contribuent aux résultats non reproductibles (voir figure 2.12). On peut remarquer que certains facteurs touchent à la structure de la communauté scientifique :

— pression à publier,
— manque d'encadrement,
— données non disponibles,
— fraude,
— relecture par les pairs insuffisante,
— manque de volonté à reproduire les études,
— manque de compétences techniques pour reproduire l'expérience ;

tandis que d'autres ont un lien avec les pratiques d'analyse des données :
— analyse de mauvaise qualité,
— manque de puissance de l'étude,
— mauvais design expérimental,

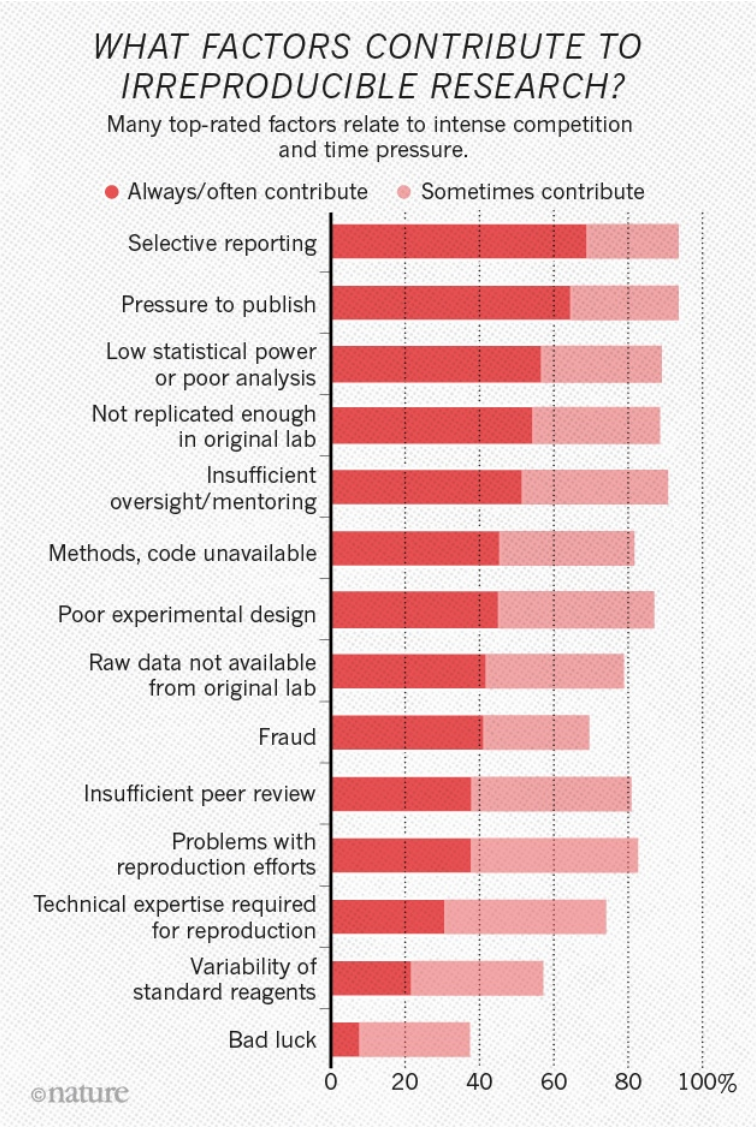


FIGURE 2.12 – **Enquête publiée dans Nature concernant la crise de reproductibilité des résultats**. Source : [Baker, 2016].

— biais de publication,
— manque de réplication en interne,
— manque de précision dans les publications (dans les méthodes, disponibilité des codes, etc.).

Peu de chercheurs considèrent que les échecs de réplication d'expériences sont dues à la variabilité des réactifs ou simplement au manque de chance.

Il semble donc admis que les pratiques d'analyses de données constituent un volet important de la crise de reproductibilité à laquelle on assiste actuellement. Et, à l'origine de ces pratiques problématiques, se trouvent des conceptions erronées à propos de certains concepts statistiques.

"*Misunderstanding or misuse of statistical inference is only one cause of the 're-*

producibility crisis', but to our community, it is an important one." [Wasserstein and Lazar, 2016]

Les concepts statistiques visiblement les plus malmenés dans la littérature scientifique sont ceux liés à l'inférence statistique et, en particulier, la P-valeur. Dans un autre article qui a fait beaucoup de bruit dans la communauté scientifique [20], Nuzzo (2014) dénonce les erreurs liées au concept de P-valeur. Elle y explique que la crise de reproductibilité en sciences a forcé la communauté scientifique à s'interroger sur la manière dont elle évaluait les résultats [Nuzzo, 2014].

Le rapport problématique des scientifiques vis-à-vis de l'inférence statistique semble être, en grande partie, un rapport problématique à la notion de P-valeur. Mais quel est exactement ce rapport à la notion de P-valeur qu'ont beaucoup de chercheurs et qui semble poser tant de problèmes ? Est-il possible de définir une praxéologie associée à la P-valeur au sein de l'institution "recherche biomédicale" ?

Pour tenter d'identifier cette praxéologie, nous allons partir de travaux d'auteurs qui se sont employés à décrire comment le concept de P-valeur était réellement utilisé dans la littérature scientifique et surtout, quels problèmes cela peut engendrer. Nous tenterons, ensuite, d'identifier la praxéologie associée à cette démarche.

2.5.4 La démarche du NHST

Dans son article, Nuzzo (2014) résume comment les théories de Fisher et de Neyman-Pearson ont été hybridées en une pratique, le *Null Hypothesis Significance Testing* (NHST).

> (...) *when UK statistician Ronald Fisher introduced the P value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense : worthy of a second look. The idea was to run an experiment, then see if the results were consistent with what random chance might produce. Researchers would first set up a 'null hypothesis' that they wanted to disprove, such as there being no correlation or no difference between two groups. Next, they would play the devil's advocate and, assuming that this null hypothesis was in fact true, calculate the chances of getting results at least as extreme as what was actually observed. This probability was the P value. The smaller it was, suggested Fisher, the greater the likelihood that the straw-man null hypothesis was false.*
>
> *For all the P value's apparent precision, Fisher intended it to be just one part of a fluid, non-numerical process that blended data and background knowledge to lead*

20. Selon la mesure *Altmetrics*, cet article est le 7e article (sur 72 542 articles) de la revue *Nature* ayant le plus fait parler de lui, et le 1er (sur 253 343 articles) de toutes les revues confondues si on le compare aux articles du même âge

> *to scientific conclusions. But it soon got swept into a movement to make evidence-based decision-making as rigorous and objective as possible. This movement was spearheaded in the late 1920s by Fisher's bitter rivals, Polish mathematician Jerzy Neyman and UK statistician Egon Pearson, who introduced an alternative framework for data analysis that included statistical power, false positives, false negatives and many other concepts now familiar from introductory statistics classes. They pointedly left out the P value.*
>
> *But while the rivals feuded — Neyman called some of Fisher's work mathematically "worse than useless"; Fisher called Neyman's approach "childish" and "horrifying [for] intellectual freedom in the west" — other researchers lost patience and began to write statistics manuals for working scientists. And because many of the authors were non-statisticians without a thorough understanding of either approach, they created a hybrid system that crammed Fisher's easy-to-calculate P value into Neyman and Pearson's reassuringly rigorous rule-based system. This is when a P value of 0.05 became enshrined as 'statistically significant', for example. "The P value was never meant to be used the way it's used today," says Goodman.*
>
> [Nuzzo, 2014]

A l'occasion de la crise de reproductibilité des résultats scientifiques, la question de la manière dont la P-valeur est utilisée refait surface. Mais elle est loin d'être récente et originale. Depuis que le concept de P-valeur a été importé dans les pratiques de recherche scientifique, des auteurs se sont élevés pour dénoncer les dérives potentielles ou avérées.

Dans ce qui suit, nous allons présenter les points de vue de trois auteurs, Carver, Cohen et Gigerenzer, qui nous permettront de donner, ensuite, une définition du NHST.

Selon Carver

> "*Statistical significance testing has involved more fantasy than fact. The emphasis on statistical significance over scientific significance in educational research represents a corrupt form of the scientific method. Educational research would be better off if it stopped testing its results for statistical significance.*" [Carver, 1978]

Carver s'intéresse à l'utilisation qui est faite de la significativité statistique (mesurée à travers la P-valeur) en recherche, dans le domaine des sciences de l'éducation.

Son message est clair : cette utilisation massive repose sur une mauvaise compréhension de ce qu'apporte réellement la P-valeur. Cette pratique corrompt la méthode scientifique et il faudrait l'arrêter au plus vite.

Carver identifie trois principales conceptions erronées à propos de ce que représente la P-

valeur :

1. **Odds-against-chance fantasy** : revient à interpréter la P-valeur comme la probabilité que les résultats soient dus à la chance. Cela revient en fait à confondre : $P(d(X) > d(x_o)|H_0)$ (P-valeur) avec $P(H_0|dx_o)$ (probabilité *a posteriori* de l'hypothèse nulle au vu des données observées).

   Il repère notamment cette conception erronée dans l'extrait suivant :

   > *"To say that the difference between two means is significant at the .01 level indicates that we can conclude, with only one chance out of 100 of being wrong, that a difference in the obtained direction would be found if we tested the whole population from which our samples were drawn"* (cité dans [Carver, 1978]).

2. **Replicability or reliability fantasy**. Une autre croyance à propos de la P-valeur est qu'elle est d'autant plus petite que les résultats ont de chance d'être reproduits. Cela revient à considérer que $1 - P$ est la probabilité de répliquer les résultats observés, autrement dit cela revient à confondre $P(d(X) > d(x_o)|H_0)$ (la P-valeur) avec $P(d(X) < d(x_s)|d(x_o)$ (la probabilité de ne pas obtenir un résultat statistiquement significatif (dépassant le seuil de significativité, $d(x_s)$) au vu des résultats observés dans l'expérience.

   Voici un exemple de texte dans lequel transparait cette croyance :

   > *"If the statistical significance is at the .05 level, it is more informative to talk about the statistical confidence as being at the .95 level. This means that the investigator can be confident with odds of 95 out of 100 that the observed difference will hold up in future investigations"* (cité dans [Carver, 1978]).

3. **Valid research hypothesis fantasy**. La dernière croyance qu'il décrit est, selon lui, la plus sérieuse. Elle consiste à interpréter $1-P$ comme étant la probabilité que l'hypothèse de recherche soit vraie compte tenu des données. Dans ce cas-ci, cela revient à tirer des conclusions à propos de l'hypothèse de recherche à partir de la probabilité d'obtenir des résultats au moins aussi extrêmes sous l'hypothèse nulle. A nouveau, rien ne nous autorise à tirer de telles conclusions.

   Des traces de pareils raisonnement sont difficiles à trouver car, selon Carver, ils sont le plus souvent implicites.

Carver identifie plusieurs raisons expliquant la persistance de ces pratiques alors qu'elles reposent sur une compréhension erronée :

1. Elle permet de défendre des résultats qui auraient été obtenus sur des petits échantillons, or les expériences sur de tels échantillons sont fréquentes en recherche en éducation et en psychologie ;
2. Elle donne un air objectif aux conclusions ;

3. Elle semble fournir une estimation du niveau de réplicabilité garanti à l'issue d'une expérience, or le caractère réplicable d'une expérience est fondamental en sciences ;
4. Elle semble fournir une mesure de l'importance d'un effet, affranchissant ainsi le scientifique de l'exercice délicat qui consiste à déterminer l'importance du résultat obtenu.

> "*Because researchers often are not able to determine whether a difference is significant or not, they are too willing to let statistics provide an objective and automatic solution, even though it is inappropriate.* " [Carver, 1978]

Pour Carver, les conceptions erronées concernant ces outils d'inférence statistique sont tellement courantes qu'elles corrompent la démarche scientifique. Il défend l'idée que la recherche en éducation se porterait mieux si on évitait purement et simplement d'utiliser ces outils. L'auteur fait référence aux outils d'inférence statistique fréquentistes (P-valeur, intervalle de confiance et test d'hypothèses) mais il ne défend pas, pour autant, le passage à des outils bayésiens.

Il évoque différentes pistes pour améliorer la méthode d'analyse des données :

1. Faire l'effort d'énoncer l'hypothèse de recherche et de l'utiliser pour prédire la direction et, si possible, l'intensité de l'effet attendu sous cette hypothèse ;
2. Mettre en œuvre des designs expérimentaux permettant de confronter efficacement cette hypothèse aux observations ;
3. Redonner plus de place à l'analyse descriptive ;
4. Se passer de P-valeurs pour se tourner, si besoin, vers d'autres outils permettant la mesure de l'intensité de l'effet, tels que l'intervalle de confiance par exemple.

Selon Cohen

Dans le domaine de la psychologie, Cohen a longtemps dénoncé les absurdités auxquelles le recours systématique à une démarche inférentielle hybride a pu mener.

> "*A colleague approaches me with a statistical problem. He believes that a generally rare disease does not exist at all in a given population, hence Ho : P = 0. He draws a more or less random sample of 30 cases from this population and finds that one of the cases has the disease, hence Ps = 1/30 = .033. He is not sure how to test Ho, chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it ? And would you believe that if he tried to publish this result without a significance test, one or more reviewers might complain ? It could happen.*" [Cohen, 1994]

Une caractéristique du NHST, selon Cohen, est l'utilisation exclusive d'hypothèses impliquant le néant (*Nil hypotheses*), l'absence totale d'effet, là où Fisher proposait une hypothèse

nulle dans le sens de *to be nullified*, à falsifier.

Le premier problème avec cette manière de définir l'hypothèse nulle, c'est que la démarche de corroboration/réfutation de l'hypothèse de recherche se transforme en une sorte de démonstration par l'absurde. Or, comme le pointe Cohen, le raisonnement de la démonstration par l'absurde (*modus tollens*) n'est plus valide dès lors que les hypothèses en jeu sont probabilistes [Cohen, 1994].

> "(...) *the "illusion of probabilistic proof by contradiction (...) [is] part of the 'hybrid logic' of contemporary statistical inference—a mishmash of Fisher and Neyman-Pearson, with invalid Bayesian interpretation.*" [Cohen, 1994]

Le second problème est que cela revient à tester une hypothèse qui, en réalité, n'est probablement jamais vraie.

> "*It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it ?*" [Cohen, 1994]

Il existera (quasiment) toujours un effet, même minime, et dès lors il sera (quasiment) toujours possible de faire une expérience qui, avec un nombre de sujet suffisant, parviendra à prouver que H_0 est fausse. Au final, savoir qu'un résultat contredit H_0 n'apporte rien sinon une indication de la direction de l'effet.

Au final, le NHST n'apporte pas grand chose or le chercheur aimerait bien que cette démarche lui permette de connaitre la probabilité qu'il se trompe. Ce décalage entre ce que le NHST apporte et ce que le chercheur voudrait qu'il apporte est responsable, selon Cohen, de la confusion qui existe entre la P-valeur et la probabilité *a posteriori* de l'hypothèse nulle.

Pour Cohen (1994), améliorer la qualité de la recherche scientifique passera par :

1. L'abandon de la recherche d'une méthode unique au profit du développement d'une boite à outils statistiques comprenant, entre autres, des outils fréquentistes (P-valeur, puissance), bayésiens, des mesures d'intensité d'effet (*effect size*) et des intervalles de confiance) ;
2. Une plus grande emphase sur la compréhension de ce que représentent les données ;
3. L'énoncé de l'hypothèse de recherche (et si test statistique il y a, ils doivent être basés sur celle-ci et pas sur l'hypothèse d'absence d'effet) ;
4. Un retour à une exigence de réplication plus systématique.

Selon Gigerenzer

Gigerenzer s'intéresse aux pratiques d'inférence statistique dans le domaine des sciences sociales. Il décrit le NHST comme un rituel, *the null ritual*, né de l'hybridation des théories de Fisher et de Neyman-Pearson et qui, comme tout rituel, devient une pratique mise en œuvre de manière automatique et non critique.

Il résume la démarche de Fisher de la manière suivante :

1. *Set up a statistical null hypothesis. The null need not be a nil hypothesis (e.g., zero difference).*
2. *Report the exact level of significance (e.g., p = .055 or .045). Do not use a conventional 5% level all the time.*
3. *Use this procedure only if you know little about the problem at hand.*

[Gigerenzer and Marewski, 2015]

Il résume la démarche de Neyman et Pearson de la façon suivante :

1. *Set up two statistical hypotheses, H1 and H2, and decide on alpha, beta, and the sample size before the experiment, based on subjective cost-benefit considerations.*
2. *If the data fall into the rejection region of H1, accept H2; otherwise accept H1.*
3. *The usefulness of this procedure is limited among others to situations where there is a disjunction of hypotheses (e.g., either $\mu 1$ or $\mu 2$ is true), where there is repeated sampling, and where you can make meaningful cost-benefit trade-offs for choosing alpha and beta.*

[Gigerenzer and Marewski, 2015]

Tandis que le NHST consisterait en la démarche suivante :

1. *Set up a null hypothesis of "no mean difference" or "zero correlation." Do not specify the predictions of your own research hypothesis.*
2. *Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < .05$, $p < .01$, or $p < .001$, whichever comes next to the obtained p value.*
3. *Always perform this procedure.*

[Gigerenzer and Marewski, 2015]

"*Now it is clear that the null ritual is a hybrid of the two theories. The first step of the ritual, to set up only one statistical hypothesis (the null), stems from Fisher's theory, except that the null always means "chance," such as a zero difference. This first step is inconsistent with Neyman–Pearson theory ; it does not specify an alternative statistical hypotheses, α, β, or the sample size. The second step, making a yes–no decision, is consistent with Neyman–Pearson theory, except that the level should not be fixed by convention but by thinking about α, β, and the sample size. Fisher (1955) and many statisticians after him (see Perlman and Wu, 1999), in contrast, argued that unlike in quality control, yes–no decisions have little role in science ; rather, scientists should communicate the exact level of significance. The third step of the null ritual is unique in statistical theory. If Fisher and Neyman–Pearson agreed on anything, it was that statistics should never be used mechanically.*" [Gigerenzer, 2004]

Selon Gigerenzer, ce rituel serait, à l'origine, une invention d'auteurs de livres vulgarisant les méthodes statistiques pour les étudiants et chercheurs des sciences sociales, invention soutenue par les éditeurs de revues scientifiques.

"*The null ritual is an invention of statistical textbook writers in the social sciences. They became familiar with Fisher's work first, mainly through his 1935 book, and only later with Neyman-Pearson theory. After learning about Neyman-Pearson, these writers (who were mostly nonstatisticians) had a problem : How should they deal with conflicting methods ? The solution would have been to present a toolbox of different approaches, but Guilford (1942), Nunnally (1975), and many others mixed the concepts and presented the muddle as a single, universal method. Indeed, the inference revolution was not led by the leading scientists. It was spearheaded by humble nonstatisticians who composed statistical textbooks for education, psychology, and other fields and by the editors of journals who found in "significance" a simple, "objective" criterion for deciding whether or not to accept a manuscript*". [Gigerenzer and Marewski, 2015]

Cette pratique se serait d'abord installée en psychologie dans les années 50 avant de se répandre à d'autres disciplines : sciences sociales, médecine, biologie, économie, sociologie, écologie [Gigerenzer, 2004].

Comment expliquer la diffusion aussi importante d'une pratique aussi éloignée des théories initiales ? Selon Gigerenzer, si les auteurs des livres de référence qui ont hybridé les deux théories portent une part de responsabilité, ils ne sont pas les seuls à blâmer. Il pointe notamment la responsabilité des chercheurs qui ne se sont pas intéressés aux méthodes utilisées, des autorités académiques qui favorisent la quantité de publication à la qualité et les éditeurs qui encouragent

ces pratiques.

> "*I once visited a distinguished statistical textbook author, whose book went through many editions, and whose name does not matter. His textbook represents the relative best in the social sciences. He was not a statistician ; otherwise, his text would likely not have been used in a psychology class. (...) . I asked the author why he removed the chapter on Bayes (...) [and] "What made you present statistics as if it had only a single hammer, rather than a toolbox ? Why did you mix Fisher's and Neyman–Pearson's theories into an inconsistent hybrid that every decent statistician would reject ?*"
>
> "(...) *There were three culprits : his fellow researchers, the university administration, and his publisher. Most researchers, he argued, are not really interested in statistical thinking, but only in how to get their papers published. The administration at his university promoted researchers according to the number of their publications, which reinforced the researchers' attitude. And he passed on the responsibility to his publisher, who demanded a single-recipe cookbook. No controversies, please. His publisher had forced him to take out the chapter on Bayes as well as the sentence that named alternative theories, he explained. At the end of our conversation, I asked him what kind of statistical theory he himself believed in. 'Deep in my heart,' he confessed, 'I am a Bayesian.'* "
>
> [Gigerenzer, 2004]

Selon cet auteur, la solution à mettre en œuvre ne consiste pas à changer de rituel (par exemple en remplaçant le NHST par une alternative bayésienne) mais à se défaire de l'idée selon laquelle il existerait une méthode qui convienne à tous les problèmes et qui peut être appliquée sans réfléchir.

2.5.5 Esquisse d'une praxéologie (E)

La définition d'une praxéologie correspondant aux pratiques d'inférence statistique est un exercice périlleux pour au moins deux raisons. Premièrement, il existe une grande diversité de problèmes en sciences et une grande diversité de pratiques liées aux tests statistiques. Deuxièmement, là où dans le savoir "savant " ou dans le savoir enseigné les démarches sont expliquées, justifiées dans un texte permettant d'identifier une praxéologie sous-jacente, dans la littérature scientifique, les raisonnements suivis en matière d'analyse de données ne sont pas explicites. Ceci explique que nous n'avons pas ici la prétention de définir *la* praxéologie liée à l'application des tests statistiques dans le milieu de la recherche biomédicale mais plutôt l'objectif d'esquisser *une* praxéologie qui, si elle n'est la seule et si elle est rarement explicite, est suffisamment courante et problématique pour être considérée comme une des causes de la crise de

reproductibilité des résultats en sciences.

Cette praxéologie liée au NHST, que nous nommerons la praxéologie E, se présente de la manière suivante.

Commençons par identifier la **technique**. Celle-ci pourrait être :

1. Partir d'une hypothèse implicite, stipulant l'existence d'un effet et sa direction, H.
2. Récolter des observations pour valider H ;
3. Poser H_0 : l'effet étudié n'existe pas ;
4. Choisir $d(X)$;
5. Modéliser $f(d(X)|H_0)$; L'hypothèse H étant imprécise (elle ne spécifie pas l'intensité de l'effet attendu), elle ne peut servir de base à la construction d'un modèle statistique ;
6. Calculer $P(d(X) > d(x_o)|H_0)$. Mesurer la cohérence entre les observations et H_0 à travers la P-valeur. Cette P-valeur peut être calculée de manière uni-directionnelle ou bi-directionnelle.
7. Tirer la conclusion de manière automatique :
   — si $P \leq 0.05$, les observations rejettent H_0 (et prouvent H),
   — si $P > 0.05$, elles prouvent H_0 si la puissance est importante et ne prouvent rien si la puissance est faible (la plupart du temps).

Cette technique semble apporter une réponse à plusieurs problèmes, à plusieurs questions différentes :

1. Comment, de manière objective, mettre un effet en évidence ?
2. Comment mesurer la pertinence scientifique d'un résultat ?
3. Comment favoriser la publication d'un article scientifique ?

Le **type de tâche** sera tantôt de déterminer de manière objective si, oui ou non, on peut affirmer qu'un certain effet expérimental a été observé, tantôt de donner une mesure de la pertinence du résultat obtenu, tantôt d'emballer une recherche dans un rituel inutile pour la science mais nécessaire pour la publication et tantôt, une combinaison de ces différentes raisons.

Cette diversité dans les tâches associées à la praxéologie E explique en partie l'importance de sa diffusion. Elle explique aussi que les **technologies** justifiant cette technique peuvent également varier d'un contexte à l'autre.

En l'occurrence, la croyance que celle-ci constitue un équivalent probabiliste à la démonstration par l'absurde peut justifier l'utilisation de cette technique pour démontrer l'existence d'un effet. Il s'agit d'une conception erronée dénoncée par Gigerenzer qui parle de "*illusion of*

probabilistic proof by contradiction" ou encore par Carver qui fait référence à la "*valid research hypothesis fantasy*".

De même, la confusion entre l'effet statistiquement significatif et l'effet substantiel ou cliniquement important justifie vraisemblablement l'utilisation de cette technique.

Enfin, il existe aussi des raisons d'utiliser la technique qui ne reposent pas sur ce que l'on pense que cette technique permet de faire, d'apporter, mais plutôt sur des considérations pragmatiques du type "sans P-valeurs, l'article aura moins de chances d'être accepté".

En définitive, ce qui justifie l'utilisation de cette technique serait soit une conception erronée à propos de ce qu'apporte réellement la P-valeur, soit une approche pragmatique voire cynique de la publication d'un travail scientifique.

Au **niveau théorique**, on peut énoncer trois éléments qui sous-tendent une ou plusieurs tâches et une ou plusieurs technologies.

Le premier élément théorique pourrait être la confiance qu'ont les scientifiques dans la littérature scientifique. Si le calcul de P-valeurs est si répandu c'est qu'il doit y avoir une bonne raison. Cet outil statistique doit probablement avoir fait ses preuves dans le domaine de la statistique théorique et doit avoir une utilité dans la littérature scientifique.

Un autre élément pourrait être l'idée que la science est une entreprise objective. Cela explique que les scientifiques cherchent à se doter d'outils objectifs permettant de dire si, oui ou non, un effet a été mis en évidence. Cela explique peut-être aussi en partie la réticence de nombreux scientifiques à accepter des outils bayésiens.

Enfin, un troisième élément qui sous-tend ce rapport au savoir est l'idée que la réfutation est un principe essentiel dans la démarche scientifique. Popper n'a-t-il pas montré qu'on pouvait prouver qu'une hypothèse était fausse mais qu'on ne pouvait jamais la valider ? Il n'est pas impossible que l'idée que la science avance par corroboration ou réfutation d'hypothèses participe à légitimer la recherche systématique de la réfutation de l'hypothèse nulle avec le NHST.

TABLE 2.9 – **Résumé de la praxéologie** E

Niveau praxéologique	**Praxéologie** E
Type de tâche	Mettre en évidence un effet en présence de bruit Mesurer la pertinence scientifique d'un résultat ? Favoriser la publication d'un article scientifique ?
Technique	1. Partir d'une hypothèse implicite, stipulant l'existence d'un effet et sa direction, H. 2. Récolter des observations pour valider H 3. Poser H_0 : l'effet étudié n'existe pas 4. Choisir $d(X)$ 5. Modéliser $f(d(X)\|H_0)$; 6. Calculer $P(d(X) > d(x_o)\|H_0)$. 7. Tirer la conclusion de manière automatique : si $P \leq 0.05$, les observations rejettent H_0 (et prouvent H), si $P > 0.05$, elles prouvent H_0 si la puissance est importante et ne prouvent rien si la puissance est faible (la plupart du temps).
Technologie	La preuve doit exister en d'autres institutions Croyances sur ce que représente la P-valeur Pragmatisme
Théorie	La science est une entreprise objective Il est possible de réfuter une hypothèse mais pas de la valider Une méthode d'analyse des données qui serait défectueuse ne se répandrait pas dans la littérature scientifique

2.5.6 Application à un cas concret

Prenons un exemple pour illustrer comment le NHST (praxéologie E) se distingue du test de significativité (praxéologie A) et du test d'hypothèses (praxéologie B).

Le contexte (librement inspiré de Kam-Hansen *et al.*, 2014) est celui du traitement de la douleur associée aux crises de migraine. La question qui intéresse le chercheur est celle de l'efficacité du traitement placebo et, en particulier de ce qui est appelé un placebo *honnête*, c'est-à-dire présenté au patient comme tel, par opposition à un placebo *classique*, présenté au patient comme un médicament potentiellement actif.

Un essai clinique est lancé en vue d'étudier l'efficacité d'un placebo honnête dans ce contexte. Des patients migraineux sont recrutés et il leur est demandé d'évaluer subjectivement leur niveau de douleur lors de deux crises de migraine, l'une non traitée et l'autre traitée par un placebo honnête. Dans cette expérience, chaque patient attribue un score de douleur pour la migraine non-traitée (NT_i) et un score de douleur pour la migraine traitée avec le placebo honnête (PH_i). La différence entre les deux scores est évaluée patient par patient : $d_i = PH_i - NT_i$. La moyenne et l'écart-type des d_i sont donnés dans le tableau 1.

Voici comment les trois démarches présentées plus haut pourraient être appliquées dans ce contexte (voir tableau 2.10).

Le **NHST** repose sur l'hypothèse implicite que le placebo honnête possède *un* effet. L'intensité attendue de cet effet n'est pas précisée seule sa direction l'est. S'agissant d'une mesure de douleur allant de 0 (douleur absente) à 10 (douleur extrême), une moyenne des différences négative signifierait par conséquent que l'intervention est efficace. L'hypothèse implicite est donc que, dans la population d'intérêt, la moyenne des différences est négative : $\delta < 0$. L'hypothèse nulle ($H_0 : \delta = 0$) est confrontée aux observations (différence moyenne $\pm$ écartype : -0.3 ± 1, $n = 64$) par l'intermédiaire de la P-valeur ($P = 0,02$). La conclusion est que le placebo honnête réduit la douleur de manière statistiquement significative ($P < 5$ %).

Le NHST utilise, on le voit, très peu d'éléments de contexte et ne nécessite pas de préciser l'hypothèse de recherche, si ce n'est la direction attendue de l'effet. En revanche, pour appliquer les deux autres démarches il nous faut développer un peu plus le contexte. On pourrait, par exemple, ajouter que des études antérieures indiquent chez ces patients une réduction de la douleur perçue de 3 points sur l'échelle de douleur pour la molécule active. L'efficacité du placebo classique est, quant à elle, estimée à la moitié de celle de la molécule active soit une diminution de 1,5 points sur cette même échelle. L'effet du placebo honnête n'est pas encore connu mais on s'attend à ce qu'il conserve une partie de l'effet du placebo classique via notamment le caractère rituel de la prise de pilule. On peut ainsi attendre que le placebo honnête induise une réduction d'environ 0,75 points sur cette échelle de douleur perçue ($H1 : \delta = -0,75$).

Sur base de ces éléments, on pourrait utiliser **le test d'hypothèses** pour déterminer la taille d'échantillon permettant de discriminer les deux hypothèses concurrentes (ici H_0 et H_1). Cela nécessiterait d'avoir une idée de la variabilité de la mesure basée sur des données d'études antérieures. Posons, par exemple, que $\sigma = 1$. Viendrait ensuite la détermination des risques α et β tolérés. Dans le présent contexte, l'erreur faite en considérant que H_1 est vraie quand, en réalité, c'est H_0 qui est vraie aurait pour conséquence de penser que le placebo honnête a un effet alors que ce n'est pas le cas. Comme implications potentielles, on pourrait imaginer le remplacement de placebos classiques par des placebos honnêtes dans de futurs essais cliniques, donc le remplacement d'une intervention efficace par une intervention inefficace. Si on estime que ce risque est sérieux, alors on pourrait juger qu'il convient de maintenir α à un niveau très bas, de l'ordre de 0,1 %, par exemple. En suivant un raisonnement du même ordre, on pourrait considérer que le risque β ne devrait pas excéder 1 %. A partir de ces éléments, on peut calculer qu'une taille d'échantillon d'un peu plus de 60 [21] individus permettrait de discriminer les deux hypothèses concurrentes en maintenant les risques α et β en dessous de 0,1 % et 1 %, respectivement.

Ces valeurs permettent de calculer *a priori* la statistique seuil à partir de laquelle on devrait considérer H_1 correcte : $Z_{seuil} = Z_{1-\alpha} = 3,09$. Si les résultats conduisent à observer une diminution de douleur perçue inférieure à $\frac{3,09\times\sqrt{n}}{S_x}$ [22], alors il faudra considérer que H_0 est correcte.

L'observation d'une différence de $0,3$ points sur l'échelle de douleur perçue (correspondant à la statistique observée : $Z_{obs} = \frac{0,3\times\sqrt{64}}{1} = 2,4 < Z_{seuil}$) doit donc nous conduire à considérer que H_0 est vraie : le placebo honnête n'aurait pas plus d'effet qu'un placebo classique.

On pourrait également utiliser le **test de significativité** pour mesurer le niveau de cohérence entre les observations et l'hypothèse initiale ($H1 : \delta = -0,75$). Partant des mêmes observations (moyenne et écart-type des différences : $= -0.3 \pm 1$, $n = 64$), on obtiendrait, cette fois-ci, une P-valeur de 0,003 indiquant que celles-ci sont très peu concordantes avec l'hypothèse initiale : l'effet du placebo honnête est donc vraisemblablement plus faible qu'attendu.

Cette comparaison permet d'illustrer deux points qui nous semblent importants.

D'une part, la conclusion à laquelle on aboutit en appliquant le NHST est sensiblement différente de celle obtenue par le test de significativité ou le test d'hypothèses. Dans le premier cas, on conclut que l'effet du placebo honnête est mis en évidence ce qui implique que l'hypothèse d'intérêt (implicite) est corroborée. Dans les deux autres cas, on conclut soit que l'effet du placebo honnête est vraisemblablement moins important qu'attendu, soit qu'au vu des risques d'erreurs préalamblement définis, il faut considérer que l'effet du placebo honnête est nul. On

21. 62 pour être précis mais nous baserons les calculs sur 64 individus pour simplifier les données numériques de cet exemple.

22. Avec S_x : l'écart-type observé dans l'échantillon. Soit $0,386$ avec 64 individus et si l'écart-type vaut 1.

TABLE 2.10 – **Comparaison de trois démarches d'inférence statistique appliquées à un même contexte.**

NHST	**Test d'hypothèses**	**Test de significativité**
	Éléments de contexte utilisés	
Direction de l'effet s'il existe $\delta < 0$	Le traitement par la molécule active réduirait la douleur de 3 points, en moyenne, contre 1,5 points pour le placebo classique. La diminution attendue avec le placebo honnête se situe, quant à elle, aux alentours de 0,75 points.	
	Hypothèses statistiques	
$H_0 : \delta = 0$ $(H_1 : \delta < 0)$	$H_0 : \delta = 0$ $H_1 : \delta = -0,75$	$H_1 : \delta = -0,75$
	Données utilisées	
$d = -0,3 \pm 1$ $n = 64$	$\sigma = 1$ (données préalables) $\alpha = 0,001$; $\beta = 0,01$ (balance coût-bénéfice) $n = 64$ (permet d'atteindre α et β) $d = -0,3 \pm 1$ (observations)	$d = -0,3 \pm 1$ $n = 64$
	Résultats	
$Z = 2,4$ $P = 0,02$	$Z_{seuil} = -3,09$ $Z_{obs} = -2,4$	$Z = -3,6$ $P = 0,003$
	Conclusions	
Le placebo honnête induit une réduction statistiquement significative de la douleur ($P < 5\%$).	Réaliser l'expérience sur une soixantaine d'individus devrait permettre de discriminer les deux hypothèses concurrentes en garantissant une faible probabilité d'erreur à long terme. Une diminution de douleur perçue inférieure à 0,386 conduit à considérer l'hypothèse nulle comme vraie : le placebo honnête ne permet pas de réduire la douleur perçue.	Le placebo honnête a peut-être un effet mais celui-ci est visiblement moins important qu'attendu.

imagine, dès lors, assez bien comment la pratique quasi-systématique du NHST peut amener les scientifiques sur de mauvaises pistes, leur donner un excès de confiance dans l'existence de certains effets et contribuer à la crise de reproductibilité en sciences.

D'autre part, le NHST ne fait pas intervenir d'éléments subjectifs dans sa mise en œuvre à l'inverse des autres démarches que nous avons décrites. En effet, dans celles-ci, une certaine subjectivité est inévitable au moment de définir l'hypothèse d'intérêt ou de déterminer la balance coût-bénéfice. D'autres chercheurs auraient très bien pu considérer que l'hypothèse d'intérêt était une diminution de la douleur perçue de 0,3 points, auquel cas le test de significativité aurait révélé que les observations sont cohérentes avec l'hypothèse initiale. De même, nous aurions très bien pu considérer que le risque α doit être fixé à 1 % au lieu de $0,1$ % ce qui nous aurait amené à devoir considérer que l'hypothèse alternative est correcte au vu des données.

Le NHST, quant à lui, ne repose que sur l'hypothèse implicite de l'existence d'*un* effet sans en préciser l'ampleur.

Notons que, dans cet exemple nous avons utilisé la P-valeur dans le cadre du NHST mais nous aurions tout aussi bien pu appliquer la même démarche à partir du risque α qui caractérise le test d'hypothèses. En plaçant automatiquement le risque α à 5 % sans considérer d'hypothèse alternative précise, nous aurions abouti au rejet de l'hypothèse H_0.

La spécificité du NHST repose donc moins sur le recours à la P-valeur que sur le caractère automatique et dé-contextualisé de la démarche. En ne faisant intervenir qu'un minimum d'éléments de contexte (uniquement le sens de la relation attendue), le NHST se place à l'abri de critiques concernant les choix à poser dans la détermination de l'hypothèse attendue notamment, en ce sens on peut dire qu'il constitue une démarche plus objective que le test de significativité ou le test d'hypothèses de Neyman et Pearson.

2.5.7 Discussion

Telle que nous l'avons énoncée, la praxéologie E, liée au NHST, semble assez proche de la praxéologie D enseignée à l'Université de Namur aux filières biomédicales. On peut noter que les techniques de D et E sont identiques alors que les types de tâches et les éléments technologico-théoriques peuvent légèrement différer. On note, entre autres, que dans la praxéologie E se retrouvent des considérations liées à l'acceptation d'un travail, d'une analyse par la communauté scientifique, ce qui n'existe pas ou peu au niveau du savoir enseigné.

Quoi qu'il en soit, il semble assez clair que la praxéologie liée au NHST diffère sensiblement de celles décrites au niveau du savoir savant. Que des praxéologies qui existent dans des institutions différentes puissent diverger n'est pas inattendu si l'on suit le raisonnement de Chevallard, au contraire. Cela doit nous inviter à rechercher les contraintes institutionnelles qui peuvent

expliquer cet écart.

En nous basant, d'une part, sur les éléments d'explication donnés par Cohen (1994), Carver (1978), Gigerenzer (2004, 2015) ou encore Nuzzo (2014) et, d'autre part, sur notre expérience du milieu de la recherche biomédicale, nous pouvons avancer l'hypothèse suivante : la recherche d'une *démonstration objective à caractère définitif* et mettant en œuvre des *méthodes standardisées* fournit un ensemble de contraintes suffisant pour expliquer la transposition institutionnelle que nous avons décrite.

1. **La recherche d'une démonstration.**

   On peut considérer que la science avance par la corroboration progressive d'hypothèses et pas par validation ou par démonstration.

   Or, au moins deux facteurs extérieurs vont pousser les scientifiques à chercher une *preuve* d'efficacité de leur traitement, à affirmer que "l'effet du traitement a été démontré" plutôt que "jusqu'à présent, les données corroborent l'hypothèse selon laquelle le traitement fonctionne".

   D'une part, le premier message est plus simple, plus clair et probablement plus facile à publier. Il est aussi probablement plus valorisé dans le cadre d'une recherche de financement.

   D'autre part, si la construction de savoir scientifique ne repose pas, en principe, sur une logique de décision (telle hypothèse est considérée vraie, tel effet a été démontré, *etc.*), il n'en reste pas moins que la recherche scientifique se déroule dans une société qui, à certains moments, implique que des décisions soient prises à propos du savoir qui se construit : considère-t-on que l'effet de la molécule a été démontré *in vitro* et que l'on peut passer à l'expérimentation animale ? Considère-t-on que tel effet est suffisamment démontré que pour investir du temps et de l'argent pour l'étudier ? Décide-t-on qu'une nouvelle molécule a effectivement prouvé son efficacité et qu'elle peut, dès lors, être mise sur le marché ?

   Certes, la praxéologie A est peut-être plus cohérente avec la logique de la démarche scientifique telle que décrite par Popper, mais il semble évident qu'il existe, dans le monde de la recherche, des contraintes qui poussent le chercheur à utiliser des outils statistiques qui semblent lui fournir la démonstration de l'efficacité d'un traitement ou, de manière plus générale, de l'existence d'un effet.

2. **La recherche d'objectivité**

   On peut considérer que la construction de savoir scientifique fait nécessairement intervenir la subjectivité du chercheur, sa capacité de discernement. Cette subjectivité intervient à différents niveaux : définition du projet de recherche, des méthodes, des mesures, du design expérimental, du nombre d'individus à inclure, de la représentation graphique des données, *etc.*

Pourquoi, dès lors, quand il s'agit de choisir la méthode d'inférence statistique ou d'interpréter le résultat de cette analyse, chercherait-on à se défaire de cette subjectivité ?

On peut avancer deux explications.

D'une part, il existe une certaine conception selon laquelle "on peut faire dire ce que l'on veut à une analyse statistique". Les scientifiques qui utiliseraient des méthodes différentes de celles communément appliquées seraient donc assez vite suspectés de l'avoir fait pour influencer la conclusion de l'analyse.

D'autre part, il y a peut-être aussi une question de facilité. Il est plus facile d'appliquer la même méthode que tout le monde que d'en développer une nouvelle et de devoir la justifier. De même, il est plus facile de laisser la P-valeur nous dire si l'effet est significatif ou non (et donc tirer, à notre place, la conclusion selon laquelle l'effet étudié a été mis en évidence) que d'utiliser notre discernement pour interpréter l'effet étudié [Carver, 1978].

Ainsi, par facilité et/ou par méfiance (peut-être en partie légitime), la démarche du NHST a pu s'installer progressivement comme la méthode d'inférence statistique de référence. Ces contraintes expliquent en partie la difficulté avec laquelle la praxéologie A aurait eu du mal à s'imposer car elle donnait une part importante à la subjectivité du scientifique, que ce soit dans la détermination de l'hypothèse à tester ou dans l'interprétation de la P-valeur.

> (...) "*no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses ; he rather gives his mind to each particular case in the light of his evidence and his ideas*".
>
> Fisher (1956), cité dans [Gigerenzer, 2004]

3. **Les obstacles à la réplication**

On peut penser que la réplication des résultats joue un rôle prépondérant dans la démarche scientifique. Pour Fisher, par exemple, une expérience ne peut, isolément, démontrer l'existence d'un phénomène.

> [The scientific researcher] *"should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation"*.
> [Fisher, 1929]

Or, différents facteurs semblent faire obstacle à la réplication des résultats en sciences et dans le domaine biomédical en particulier. Le premier est le fait que les scientifiques sont soumis à une importante pression pour publier le résultat de leurs recherches [Baker, 2016] et que les expériences répliquant purement et simplement une expérience déjà publiée ne sont généralement pas considérées comme très faciles à publier, peut-être

par manque d'originalité. A cela s'ajoute un problème éthique lié à la réplication d'expériences qui impliquent des êtres humains. En effet, de telles expériences reposent sur le principe de l'*équipoise clinique* qui ne se marie pas facilement avec l'idée de réplication des expériences [Kowalski, 2010].

Ces deux obstacles à la réplication favorisent sans doute une recherche de démonstration à caractère définitif et donc le recours à un outil d'inférence statistique tel que le NHST.

4. **La complexité de maitriser un large panel de méthodes**

   De nombreux statisticiens semblent aujourd'hui s'accorder sur l'idée que les différents outils d'inférence statistique ne s'opposent pas mais doivent être considérés comme un ensemble complémentaire et constituer une "boîte à outils" [Kass, 2011].

   L'idéal serait donc que les scientifiques aient une formation solide aux différents outils d'inférence statistique notamment. Ils sont, en effet, nombreux à reconnaitre que ce manque de formation a joué un rôle important dans la crise de reproductibilité des résultats scientifiques [Baker, 2016].

   Cependant, atteindre le niveau nécessaire à utiliser de manière avisée chacun des outils d'inférence statistique est loin d'être évident. C'est pourquoi nous pensons que le manque de formation des scientifiques dans le domaine de la statistique continuera à favoriser une certaine homogénéité dans les méthodes d'inférence statistique plutôt que la maîtrise d'une boîte à outils.

Dans le domaine des pratiques sociales de référence, ici en l'occurrence dans le domaine de la recherche scientifique, il semble possible d'identifier différentes caractéristiques du milieu qui expliquent la nature et l'intensité de la transposition institutionnelle que nous avons observée entre le savoir savant et les pratiques sociales de référence et donc, indirectement, entre le savoir savant et le savoir enseigné.

Ces caractéristiques sont les suivantes : le milieu de la recherche scientifique semble pousser les chercheurs à appliquer une méthodologie standardisée (plutôt qu'un panel de méthodes variées) qui leur permettrait d'apporter des preuves définitives (plutôt qu'une corroboration provisoire) et objectives (à l'abri des soupçons liés à l'interprétation subjective) de l'existence des effets étudiés.

Cependant, le paysage de la recherche biomédicale n'est pas simple et figé. Et si certains facteurs jouent en faveur de l'installation et de l'enracinement d'un rituel statistique dans les pratiques d'analyse des données, il en existe, heureusement, d'autres qui poussent les pratiques dans la bonne direction, c'est-à-dire dans la direction d'une utilisation adéquate et complémentaire des différents outils d'inférence statistique.

1. La **formation**, initiale et continue, des scientifiques à la démarche scientifique et à l'inférence statistique ;

2. Les **mécanismes** (auto-)**correcteurs** que sont la relecture par les pairs, l'expérience, la capacité d'analyse des problèmes liés à la recherche scientifique (par exemple la capacité à faire le point à la suite de la crise de reproductibilité des résultats scientifiques) ;

3. Les ***guidelines*** qui ont été mises au point pour éviter un certain nombre de dérives dans les pratiques d'analyses de données : citons ARRIVE (*Animal Research : Reporting of In Vivo Experiments*, [Kilkenny et al., 2010]), STROBE (*STrengthening the Reporting of OBservational studies in Epidemiology* [Vandenbroucke et al., 2007]) et CONSORT (*CONsolidated Standards Of Reporting Trials* [Moher et al., 2010]). Ces *guidelines* sont développées par des groupes d'experts, notamment en biostatistique, afin de se mettre d'accord sur les pratiques à favoriser et celles à éviter. Elles permettent, en principe, aux scientifiques et aux éditeurs de se mettre d'accord sur ce que peut être une bonne manière d'analyser les données ;

4. **L'évolution des règles** régissant la structure de la communauté scientifique. Certaines règles et exigences à propos de la recherche scientifque sont de nature à faire évoluer la qualité des analyses publiées. Citons, par exemple, l'exigence d'enregistrer les protocoles des essais cliniques avant leur démarrage qui permet de lutter contre le biais de publication, ou d'enregistrer les plans d'analyse statistique à l'avance afin de lutter contre le biais d'analyse.

5. L'impact des **méta-analyses**, qui tendent à favoriser une vue d'ensemble dans laquelle chaque expérience individuelle n'apporte pas une démonstration définitive mais contribue à estimer l'effet vraisemblable d'une intervention.

Jusqu'à présent, les contraintes qui ont pesé sur l'institution "recherche biomédicale" ont globalement favorisé l'installation et la diffusion d'une praxéologie (E) dommageable pour la construction du savoir scientifique.

Ces contraintes nous semblent pouvoir expliquer comment les praxéologies existant dans le domaine du savoir savant ont été à ce point modifiées lorsqu'elles ont été importées dans les pratiques de recherche. Dans cette institution, et au vu des contraintes actuelles, la praxéologie du NHST semble mieux répondre aux contraintes qu'une praxéologie basée sur une combinaison adéquate des différents outils d'inférence statistique.

Si ces contraintes persistent, il y a fort à parier que n'importe quel outil d'inférence statistique, qu'il soit bayésien ou fréquentiste, risque de subir une transposition institutionnelle qui pourrait pervertir l'outil de la même manière que ce que nous avons décrit ici. Ainsi, le changement d'un outil pour un autre ne nous semble pas, de ce point de vue, constituer une solution satisfaisante. C'est également le message de Gigerenzer (2015) qui met en garde contre le remplacement d'un rituel fréquentiste par un rituel bayésien.

Cependant, on peut aussi espérer que les contraintes changent à l'avenir et permettent de

mettre en place un milieu propice au développement et à la diffusion de pratiques d'inférence statistique plus adéquates.

2.6 Conclusion

La question de recherche que nous avons posée en début de chapitre était la suivante :

Quelles contraintes déterminent les diverses praxéologies existant au niveau du savoir enseigné ?

A partir de l'analyse du savoir savant nous tentons de définir trois praxéologies distinctes :

1. Une praxéologie (A) liée au test de significativité selon Fisher ;
2. Une praxéologie (B) liée au test d'hypothèses selon Neyman et Pearson ;
3. Une praxéologie (C) liée au test statistique bayésien basé sur la probabilité *a posteriori*.

L'analyse du savoir enseigné au niveau local nous permet de définir une quatrième praxéologie (D). Celle-ci ne correspond exactement à aucune des praxéologies décrites dans le savoir savant et semble plutôt hybrider des caractéristiques des praxéologies A, B et, dans une moindre mesure, C.

Les contraintes qui agissent, en général, lors de la *mise en texte* d'un savoir mathématique nous semblent favoriser l'enseignement d'une praxéologie qui repose sur une démarche débarrassée de tout élément subjectif. Ceux-ci sont, en effet, difficilement mis en texte, il est difficile d'en programmer l'apprentissage, de la rendre publique et d'en évaluer l'acquisition.

De plus, les contraintes spécifiques à notre niveau local (volume horaire, profil de l'équipe enseignante, visée appliquée de l'enseignement) favorisent une praxéologie qui repose sur un seul outil d'inférence plutôt qu'une boîte à outils et, surtout, qui montre une proximité suffisante avec ce que l'on observe au niveau des pratiques sociales de référence.

Ce constat nous amène à nous intéresser de près à ces pratiques sociales de référence.

Un rapide tour d'horizon révèle que les principaux outils d'inférence statistique que l'on retrouve dans la littérature médicale sont la P-valeur, l'intervalle de confiance, dans une moindre mesure la notion de puissance et plus rarement encore, les outils d'inférence bayésiens.

Il apparait également que de nombreux auteurs ont dénoncé la manière dont ces notions d'inférence statistique sont couramment utilisées par les chercheurs. L'étude des arguments soulevés par plusieurs de ces auteurs (Carver, Cohen, Gigerenzer ou encore Nuzzo) nous amène à

esquisser une praxéologie (E) qui correspond à ces pratiques dénoncées et est parfois dénommée le *Null Hypothesis Significance Testing* (NHST).

Cette praxéologie E, tout comme la praxéologie enseignée, se distingue de celles existant dans le savoir savant par le fait qu'elle peut être appliquée de manière assez automatique et objective. En effet, hormis le sens attendu de l'effet étudié, le NHST ne fait pas intervenir d'éléments de contextes, à la différence du test de significativité ou du test d'hypothèses.

Il semble que les contraintes que l'on pourrait trouver au niveau de la recherche biomédicale − pressions éditoriales, besoin de messages clairs, risques de biais dans l'interprétation subjective des résultats, difficultés éthiques liées à la reproductibilité des expériences, *etc.* − favorisent l'émergence de praxéologies reposant sur une démarche objective d'analyse de données, sur une démarche relativement standardisée, unique, et permettant d'obtenir des réponses claires et définitives. L'existence de ces contraintes permet d'expliquer le succès du NHST et les difficultés que l'on peut attendre si l'on souhaite remplacer cet outil.

Notons que si cette institution renferme des contraintes de nature à simplifier et pervertir des outils d'inférence statistique, elle en contient aussi d'autres qui tendent à complexifier, nuancer et corriger les mauvaises pratiques d'analyse de données.

En dernière analyse, nous retiendrons que les praxéologies D et E sont assez proches notamment parce qu'elles sont influencées par des facteurs semblables (la recherche d'objectivité, la tendance à l'homogénéisation des méthodes d'analyse de données) mais aussi probablement parce qu'elles s'influencent mutuellement, entretenant ainsi un cercle vicieux bien énoncé par Cobb (2014, cité dans [Wasserstein and Lazar, 2016]) :

> "*Q[question] : Why do so many colleges and grad schools teach p = 0.05 ?*
> *A[answer] : Because that's still what the scientific community and journal editors use.*
> *Q : Why do so many people use p=0.05 ?*
> *A : Because that's what they were taught in college or grad school.*"

Chapitre 3

Ingénierie didactique

Partant de difficultés d'enseignement locales et des conséquences potentiellement importantes que celles-ci peuvent engendrer au niveau de la recherche biomédicale nous avons, dans l'introduction, posé deux questions principales.

Dans le chapitre précédent, nous interrogeons le savoir enseigné : qu'enseigne-t-on ? Quels écarts avec le savoir "savant" et avec les pratiques sociales de référence ? Quelles contraintes s'appliquent au savoir enseigné et quels types de praxéologies favorisent-elles ?

En parallèle à cette réflexion [1], nous tenterons, dans le présent chapitre, d'interroger le *comment* : comment enseigner l'inférence statistique aux étudiants des filières biomédicales à l'Université de Namur ? Le recours à une situation fondamentale au sens de Brousseau constitue-t-il une piste pertinente pour l'enseignement de l'inférence statistique ?

Pour y répondre, nous suivrons la démarche suivante.

Partant du principe que le savoir à enseigner est constitué principalement par le test d'hypothèses, nous mettrons au point un dispositif expérimental censé permettre aux étudiants l'apprentissage de la démarche du test d'hypothèses selon Neyman et Pearson.

Nous l'expérimenterons sur notre public cible et tenterons, d'une part, d'évaluer la pertinence de ce dispositif et, d'autre part, de décrire les principaux obstacles à cet apprentissage.

Avant de mettre en œuvre cette démarche, il nous faudra, au préalable, présenter brièvement le cadre théorique sur lequel nous nous appuyons. En effet, nous utiliserons ici le cadre conceptuel de la théorie des situations didactiques de Brousseau qui nous semble le plus adapté à l'objectif que nous poursuivons dans ce chapitre. Pour citer Chevallard :

" [La théorie des situations] *tend à privilégier le point de vue de l'*économie *et à*

1. Les deux chapitres ont été construits en parallèle si bien qu'il n'est pas possible de présenter l'un comme étant la conséquence logique de l'autre.

> *laisser un peu en retrait le point de vue de l'*écologie *des systèmes. Ou, pour le dire plus concrètement, elle tend à se centrer sur le* fonctionnement *de la machine, en laissant un peu de côté l'étude des* conditions de possibilité *de ce fonctionnement [...] Ou encore, pour avancer une autre formule, plus personnalisée mais au demeurant tout aussi approximative que les précédentes : Guy Brousseau me paraît "obsédé" par les conditions du* bon *fonctionnement des systèmes didactiques ; je suis, quant à moi, davantage fasciné par l'étude des conditions de possibilité de leur fonctionnement tout court − bon ou moins bon.*" (p.103) (Chevallard (1992) cité dans [Margolinas, 2005]).

La théorie des situations didactiques nous fournira donc des éléments théoriques nous permettant de penser notre dispositif d'enseignement expérimental ainsi qu'un cadre pour analyser les difficultés des étudiants en termes d'obstacles. L'ingénierie didactique telle que présentée par Artigue constituera la méthodologie avec laquelle nous tenterons de répondre à la question de recherche.

3.1 Cadre théorique et question de recherche

3.1.1 Théorie des situations didactiques

La théorie des situations didactiques (TSD) représente une des principales théories en didactique des mathématiques, elle est due à Brousseau et voit le jour durant les années 1970 et 1980 [Margolinas, 2005].

Dans cette section nous allons présenter brièvement les éléments de ce cadre conceptuel qui nous sont apparus essentiels.

De manière très générale, on peut définir la TSD comme une théorie **constructiviste** s'intéressant à certains phénomènes liés à l'apprentissage dans un contexte où il existe une intention d'enseigner. Cette théorie s'est construite en opposition aux conceptions *transmissives* et *comportementalistes* (*béhaviouristes*) de l'apprentissage.

Ainsi, la conception *transmissive* est celle selon laquelle :

> (...) *l'élève est supposé totalement ignorant de l'objet d'apprentissage tandis que le maître sait, et selon laquelle le rôle de l'enseignant consiste à déverser le savoir dans une "tête vide"* [Régnier, 2012].

Selon cette conception, enseigner consiste à découper la matière en éléments suffisamment petits et clairs pour être transmis efficacement. Ensuite, si l'élève n'apprend pas, c'est qu'il n'a pas bien écouté.

La conception *comportementaliste* est celle selon laquelle :

> (...) *on ne peut avoir accès aux structures mentales du sujet mais pour laquelle on peut faire acquérir au sujet un savoir en lui aménageant un certain nombre d'étapes bien choisies et graduées* [Régnier, 2012].

Selon cette conception, enseigner consiste à découper la matière à enseigner en morceaux suffisamment petits puis à conditionner l'élève pour qu'il donne la bonne réponse au bon stimulus, la répétition permettant d'ancrer la relation stimulus-réponse.

Par opposition, le courant constructiviste décrit l'apprentissage comme l'intégration d'un savoir nouveau dans un réseau de savoir déjà formé. Pour apprendre, un élève doit "reconstruire" le savoir et l'intégrer à ce qu'il sait déjà voire transformer ses conceptions pour pouvoir accueillir le savoir nouveau.

> *"These constructivist approaches have permitted people to have a new look at learning, showing that it cannot be reduced to a simple processe of transmission of facts. What can be learned is strongly constrained by the subjects'initial conception - by the situations that are proposed to them and the means of actions that are given to them for these situations"* [Artigue, 1999].

Dans sa théorie des situations didactiques Brousseau a été fortement inspiré par les idées constructivistes, cependant il a progressivement abandonné l'idée de développer une didactique purement constructiviste. Ainsi, comme nous allons le voir, Brousseau insistera notamment sur la nécessité de phases d'*institutionnalisation* du savoir, qui se détachent du constructivisme, pour l'installation des savoirs [Brousseau, 1986].

Ainsi Brousseau se détache de ce qu'il appelle le *constructivisme radical*.

> *Le constructivisme radical est une théorie pédagogique qui affirme que l'élève ne s'approprie que les connaissances qu'il produit lui-même. Elle assure donc en d'autres termes, que sans autre intervention didactique que le choix des situations non didactiques appropriées, les élèves peuvent (doivent) produire, par une construction autonome, des connaissances équivalentes à celles que la société veut leur enseigner (et qu'elle a elle-même construites de façon non didactique)* [Brousseau, 2010].

> *"La conséquence la plus spectaculaire des études théoriques du contrat didactique a été de montrer que le constructivisme radical ne peut pas aboutir à l'acquisition des savoirs visés par l'élève sans intervention didactique"* [Brousseau, 2000].

En tant que théorie *constructiviste* de l'apprentissage, la TSD propose d'interpréter les erreurs des élèves non pas comme un manque mais plutôt comme la trace d'une connaissance déjà présente chez l'élève et qui possède un certain domaine de validité. Cette manière de

considérer les erreurs chez l'élève permet d'expliquer deux caractéristiques essentielles de la théorie développée par Brousseau : la recherche de *situations fondamentales* pour une notion à enseigner et l'analyse des difficultés en termes d'*obstacles*.

La situation fondamentale

Pour bien comprendre le concept de situation fondamentale, il convient, tout d'abord, de faire la distinction entre connaissance et savoir.

> *"Le savoir est une construction sociale, qui résulte d'un processus historique. La connaissance est un acquis personnel, signe de l'équilibre entre un sujet et son milieu"* [Margolinas, 2005].

Brousseau pose alors la question des *conditions d'émergence d'une connaissance*. Historiquement, quels ont été les contraintes nécessaires pour qu'une connaissance émerge ?

Cette question amène à s'intéresser au concept de *situation*, c'est-à-dire *l'ensemble des circonstances dans lesquelles une personne se trouve, et des relations qui l'unissent à son milieu"* [Brousseau, 1986].

Ainsi Brousseau cherche à déterminer quelles étaient les caractéristiques de la situation *mathématique* à laquelle les mathématiciens ont été confrontés lorsqu'ils ont développé la connaissance qui est ensuite devenue le savoir à enseigner. Et, partant des éléments qui ont, historiquement, justifié le développement d'une connaissance, Brousseau propose d'essayer de construire des situations *didactiques*, c'est-à-dire des situations ayant pour but de provoquer l'émergence d'une certaine connaissance chez un individu. Il s'agit donc de reconstruire, artificiellement, une situation contenant les conditions nécessaires et suffisantes d'émergence d'une connaissance, en s'inspirant de l'histoire de cette connaissance [Margolinas, 2005].

> "*Modelling a teaching situation consists of producing a game specific to the target knowledge, among different subsystems : the educational system, the student system, the milieu, etc. (...) The game must be such that the knowledge appears in the chosen form as the solution, or as the means of establishing the optimal strategy*"[Brousseau, 1997, p.47].

> "[La situation fondamentale] *offre des possibilités d'enseignement mais surtout une représentation du savoir par les problèmes où il intervient permettant de restituer le sens du savoir à enseigner"* [Brousseau, 2010].

Il distingue trois types de situations fondamentales [Brousseau, 2005] :

1. La situation fondamentale **générique** : elle permet de déterminer n'importe quelle situation fondamentale d'une discipline en faisant varier les variables didactiques qui la composent. Elle apparaît surtout utile pour organiser les connaissances ou développer des situations similaires (*isomorphes*).
2. La situation fondamentale **signifiante** : il s'agit d'une situation fondamentale d'une notion dans laquelle le sens global est conservé même s'il est justifié par des méthodes simplifiées. Les méthodes peuvent être par la suite complexifiées tout en gardant le sens global.
3. La situation fondamentale **génétique** : il s'agit d'un ensemble dynamique de situations dans lequel chaque situation engendre une connaissance qui appelle la connaissance suivante tout en nécessitant un minimum d'apports extérieurs.

Par ailleurs, les situations fondamentales peuvent être catégorisées selon l'activité qui est attendue du sujet. Brousseau (1997) distingue ainsi les situations :

1. D'**action**, dans lesquelles ce qui est attendu du sujet est une action, qui peut être basée sur des stratégies intuitives et implicites. Les règles du jeu permettront de déterminer si l'action était bonne ou mauvaise mais sans avoir besoin de justification ;
2. De **formulation**, dans lesquelles ce qui est attendu du sujet est, non seulement de choisir une action, mais de devoir identifier ses propres stratégies implicites, de les communiquer à un autre sujet et de les comparer dans le but de se mettre d'accord sur l'action à effectuer ;
3. De **validation**, dans lesquelles le but est de mettre à l'épreuve des affirmations et de produire des théorèmes.

Bien que centrales dans sa théorie, Brousseau nous dit que l'enseignement ne doit pas se réduire à la présentation de situations fondamentales. Il distingue notamment deux phases dans un processus d'enseignement : la phase *a-didactique* et la phase *didactique*.

> *The modern conception of teaching therefore requires the teacher to provoke the expected adaptation in her students by a judicious choice of "problems" that she puts before them. These problems, chosen in such a way that students can accept them, must make the students act, speak, think, and evolve by their own motivation. Between the moment the student accepts the problem as if it were her own and the moment when she produces her answer, the teacher refrains from interfering and suggesting the knowledge that she wants to see appear. The student knows very well that the problem was chosen to help her acquire a new piece of knowledge, but she must also know that this knowledge is entirely justified by the internal logic of the situation and that she can construct it without appealing to didactical reasoning.*

> *Not only can she do it, but she must do it because she will have truly acquired this knowledge only when she is able to put it to use by herself in situations which she will come across outside any teaching context and in the absence of any intentional direction. Such a situation is called an adidactical situation* [Brousseau, 1997, p.48]

La phase a-didactique est celle durant laquelle l'enseignant se retire en apparence pour laisser l'élève interagir avec la situation. Au cours de cette phase, l'enseignant tente de transférer la responsabilité de l'apprentissage vers l'apprenant lui-même, un processus que Brousseau nomme la *dévolution*. L'idée est que la situation elle-même va provoquer la modification des connaissances et non pas directement l'enseignant.

> *"The teacher must, however, accept responsibility for the results and ensure that the student has the effective means for acquiring knowledge. This "making sure" is fallacious but essential if she is to be allowed to engage the student's responsibility. Similarly, the student must accept responsibility for solving some problems whose solutions she has not been taught, although she does not see a priori the choices that are offered her and their consequences, and she is therefore involved in an obvious instance of juridical irresponsibility"* [Brousseau, 1997, p.50].

Pour parler de ce transfert, implicite, de responsabilité entre les élèves et le maître, Brousseau invoque la notion de *contrat didactique* qui détermine les responsabilités réciproques et les règles du jeu.

Ce concept permet d'expliquer certains phénomènes didactiques dont un exemple célèbre est "l'âge du capitaine". Il s'agit d'une situation simple utilisant un énoncé du type :

"Sur un bateau, il y a 26 moutons et 10 chèvres. Quel est l'âge du capitaine ?" [Astolfi, 1990].

De manière étonnante, les élèves à qui l'on pose cette question sont nombreux à donner comme réponse "36 ans". Une explication simple serait que certains élèves font "n'importe quoi" [Astolfi, 1990]. Le concept de contrat didactique permet de s'intéresser aux règles auxquelles les élèves semblent obéir lorsqu'ils donnent cette réponse. En effet, les élèves agissent comme si :

- l'enseignant ne pose pas de question absurde ;
- l'enseignant fournit aux élèves les moyens de répondre à la question ;
- les élèves doivent utiliser les données à leur disposition pour résoudre le problème ;
- la réponse donnée doit être plausible.

> *"Plus souvent qu'on ne le croit, les élèves ne répondent pas vraiment à la question posée, comme ils le feraient dans un contexte plus neutre, mais ils répondent d'abord à l'enseignant, en s'efforçant - à tort ou à raison - de décoder la nature de son attente*

et de s'y ajuster positivement. Comme dit Chevallard, ils raisonnent sous influence" [Astolfi, 1990].

Durant la phase a-didactique, l'élève développe des connaissances implicites, tente de communiquer ses stratégies en utilisant certains concepts, se fait une représentation de l'utilité de cette nouvelle connaissance. Mais ensuite, au cours de la phase d'*institutionnalisation*, l'enseignant doit idéalement reprendre la main pour désigner aux élèves le savoir auquel les connaissances nouvellement formées font référence, expliquer à quoi il peut servir en dehors des cas qui ont été explorés et donner le vocabulaire socialement admis [Brousseau, 1997].

L'organisation de l'enseignement en phases a-didactiques et didactiques requiert énormément de temps. En effet, le sujet doit avoir le temps de se familiariser avec le problème, de tester des stratégies implicites, de les énoncer, il faut ensuite les valider et institutionnaliser les connaissances qui auront été produites. L'application pratique des éléments issus de la théorie des situations didactiques peut donc sembler particulièrement complexe voire impossible. En réalité, la question n'est pas tellement de savoir *s'il faut* organiser tout un enseignement sur base de ces principes mais plutôt de déterminer *à quelles notions*, cruciales au sein d'une discipline, réserver ce type d'approche.

A ce propos, Brousseau propose que l'on réserve cette approche aux notions les plus importantes d'une discipline et qui semblent résister à l'enseignement.

A partir des éléments précédents, il nous est possible de proposer une définition du concept de *situation fondamentale* de Brousseau. Il s'agit d'une situation :

— a-didactique ;
— qui comporte les conditions d'émergence de la connaissance visée ;
— pensée de manière à rendre la connaissance visée plus efficace que les autres connaissances.

La notion d'obstacle

L'obstacle épistémologique

Dans son livre "La formation de l'esprit scientifique" (1934), Bachelard décrit les errements des penseurs des époques pré-scientifiques (XVIe, XVIIe, XVIIIe siècles) et tente de catégoriser les obstacles auxquels ces penseurs sont confrontés de manière récurrente et qui les empêchent d'atteindre une connaissance scientifique.

Selon lui, pour trouver les conditions psychologiques des progrès scientifiques, il est intéressant de penser *en terme d'obstacles*. De plus, il convient de ne pas se limiter aux obstacles externes (la complexité des phénomènes étudiés) ou internes (les limites humaines) mais plutôt d'investiguer les obstacles propres à l'acte d'apprendre.

Ainsi il décrira différents obstacles qu'il illustre par de nombreux exemples tirés de textes pré-scientifiques. Parmi eux, on peut citer l'obstacle de l'expérience première, la connaissance générale, l'obstacle verbal, la connaissance unitaire et pragmatique, l'obstacle substantialiste, l'obstacle réaliste, l'obstacle animiste ou encore l'obstacle de la connaissance quantitative [Bachelard, 1934].

Ces obstacles peuvent avoir pour origine des éléments de l'ordre de l'instinct (les connaissances usuelles, les opinions ou encore la valorisation des idées fréquemment utilisées) ou, au contraire, les observations empiriques.

A propos des connaissances usuelles, il considère que l'on apprend en regardant les erreurs faites avec la connaissance antérieure, et affirme que "*on connait contre une connaissance antérieure*" [Bachelard, 1934].

> *"Face au savoir, ce qu'on croit savoir clairement offusque ce qu'on devrait savoir. Quand il se présente à la culture scientifique, l'esprit n'est jamais jeune. Il est même très vieux, car il a l'âge de ses préjugés"* [Bachelard, 1934].

Mais les connaissances usuelles ne sont pas les seules à être à l'origine d'obstacles : c'est également le cas des opinions, dont il faut, selon lui, se défaire si l'on veut accéder à une connaissance scientifique. Et même lorsque la connaissance est acquise par un *effort scientifique*, elle peut décliner. Avec l'âge, notre esprit utilise les idées et valorise (indûment) celles qui lui servent le plus souvent. Cela peut constituer une source d'obstacles.

Si les obstacles peuvent avoir pour origine les instincts, les observations empiriques peuvent, paradoxalement, en être aussi la cause. En effet, en voulant rationaliser une connaissance empirique, on risque de la dénaturer car :

> *"D'une manière bien visible, on peut reconnaitre que l'idée scientifique trop familière se charge d'un concret psychologique trop lourd, qu'elle amasse trop d'analogies,*

> *d'images, de métaphores, et qu'elle perd peu à peu son vecteur d'abstraction, sa fine pointe abstraite. Si bien que l'on peut remettre en question l'idée que le savoir sert automatiquement au savoir* [Bachelard, 1934].

Pour le didacticien, l'intérêt de cette analyse des obstacles épistémologiques est qu'elle fournit une hypothèse pour comprendre les problèmes d'apprentissage rencontrés par les élèves. D'ailleurs, l'auteur fait le parallèle entre les obstacles épistémologiques et les obstacles pédagogiques.

L'obstacle didactique

Dans la théorie des situations didactiques, Brousseau s'inspire du concept d'obstacle épistémologique de Bachelard. Il faut, toutefois, noter deux différences majeures.

Premièrement, Brousseau s'intéresse aux errements observés chez les élèves et non plus chez les penseurs pré-scientifiques. Il parlera donc d'obstacles *didactiques* car ils surviennent dans un contexte d'apprentissage scolaire.

Deuxièmement, Brousseau considère les obstacles dans le domaine des mathématiques, alors que Bachelard restait dans le domaine des sciences empiriques.

> *"L'erreur n'est pas seulement l'effet de l'ignorance, de l'incertitude, du hasard que l'on croit dans les théories empiristes ou béhavioristes de l'apprentissage, mais l'effet d'une connaissance antérieure, qui avait son intérêt, ses succès, mais qui, maintenant, se révèle fausse, ou simplement inadaptée. Les erreurs de ce type ne sont pas erratiques et imprévisibles, elles sont constituées en obstacles. Aussi bien dans le fonctionnement du maître que dans celui de l'élève, l'erreur est constitutive du sens de la connaissance acquise"* [Brousseau, 1998].

Ainsi, plutôt que d'étudier les erreurs commises par les élèves, l'analyse en termes d'obstacles invite à chercher ce que les élèves savent, qui possède un domaine de validité et qui les empêche de construire la connaissance visée.

> *"Un obstacle se manifeste donc par des erreurs, mais ces erreurs ne sont pas dues au hasard. Fugaces, erratiques, elles sont reproductibles, persistantes. De plus, ces erreurs, chez un même sujet, sont liées entre elles par une source commune : une manière de connaître, une conception caractéristique, cohérente sinon correcte, une " connaissance" ancienne et qui a réussi dans tout un domaine d'actions"* [Brousseau, 1998].

Un obstacle est donc une connaissance qui possède un domaine de validité, qui "résiste et reparait" et qui se manifeste par des erreurs.

"L'obstacle est constitué comme une connaissance, avec des objets, des relations, des méthodes d'appréhension, des prévisions, avec des évidences, des conséquences oubliées, des ramifications imprévues... Il va résister au rejet, il tentera comme il se doit, de s'adapter localement, de se modifier aux moindres frais, de s'optimiser sur un champ réduit, suivant un processus d'accommodation bien connu.

C'est pourquoi, il faut un flux suffisant de situations nouvelles, inassimilables par lui, qui vont le déstabiliser, le rendre inefficace, inutile, faux, qui vont en rendre nécessaire la reprise ou le rejet, l'oubli, la scotomisation — jusque dans ses ultimes manifestations.

Aussi, le franchissement d'un obstacle exige un travail de même nature que la mise en place d'une connaissance, c'est-à-dire des interactions répétées, dialectiques de l'élève avec l'objet de sa connaissance" [Brousseau, 1998].

Une connaissance, comme un obstacle, est toujours le fruit d'une interaction de l'élève avec son milieu et plus précisément avec une situation qui rend cette connaissance "intéressante" " [Brousseau, 1998]

Brousseau distingue trois origines possibles pour les obstacles dans le contexte didactique [Brousseau, 1998].

1. **Ontogénique** si l'obstacle est lié aux limites de l'élève à un moment de son développement ;
2. **Didactique** si l'obstacle "[semble] *ne dépendre que d'un choix ou d'un projet du système éducatif*" ;
3. **Épistémologique** si l'obstacle est *"constitutif de la connaissance visée"*.

"Les obstacles d'origine proprement épistémologique sont ceux auxquels on ne peut, ni ne doit échapper, du fait même de leur rôle constitutif dans la connaissance visée. On peut les retrouver dans l'histoire des concepts eux-mêmes. Cela ne veut pas dire qu'on doit amplifier leur effet ni qu'on doit reproduire en milieu scolaire les conditions historiques où on les a vaincus" [Brousseau, 1998].

Montrer que des erreurs sont dues à un obstacle épistémologique est une entreprise complexe. Il faut tout d'abord montrer qu'il s'agit d'un obstacle et que celui-ci ne dépend ni des caractéristiques de l'élève à un moment de son développement ni des choix didactiques qui ont été posés mais qu'il est constitutif du savoir. Ensuite, il faut trouver la trace de ces obstacles à la fois dans l'histoire des mathématiques et dans les modèles spontanés des élèves et trouver les conditions pédagogiques du franchissement de ces obstacles [Brousseau, 1998].

3.1.2 Ingénierie didactique

L'ingénierie didactique désigne à la fois un dispositif d'enseignement et une méthodologie de recherche [Artigue, 1988]. Dans ce qui suit, nous nous focaliserons sur l'ingénierie didactique en tant que méthodologie de recherche.

Celle-ci a pour objectif général l'élaboration des connaissances didactiques par la confrontation entre ce que prédit la théorie didactique et ce qui est observé. Cette méthodologie expérimentale se distingue des autres (enquêtes, questionnaires, interviews, etc.) de plusieurs façons.

D'une part, ces expérimentations ont lieu **en situation de classe**, afin de capturer au mieux l'ensemble des phénomènes liés à l'apprentissage en situation scolaire. Ces expériences suivent le schéma suivant : "*conception, réalisation, observation et analyse de séquences d'enseignement*" [Artigue, 1988].

D'autre part, par rapport à d'autres types de recherches ayant lieu en situation de classe, l'ingénierie didactique utilise une **validation interne**, *fondée sur la confrontation entre analyse a priori et analyse a posteriori* et non pas sur une validation *externe* qui utiliserait la comparaison statistique entre un groupe expérimental et un groupe témoin [Artigue, 1988].

Artigue distingue différentes phases successives dans la méthodologie de l'ingénierie didactique :

La première phase, **l'analyse préalable**, repose sur un cadre didactique général et contient, généralement les analyses suivantes :

- *"l'analyse épistémologique des contenus visés par l'enseignement* ;
- *l'analyse de l'enseignement usuel et de ses effets* ;
- *l'analyse des conceptions des élèves, des difficultés et obstacles qui marquent leur évolution* ;
- *l'analyse du champ de contraintes dans lequel va se situer la réalisation didactique effective"* [Artigue, 1988].

Ces analyses permettent de décrire en quoi la situation actuelle est insatisfaisante et laissent entrevoir les variables sur lesquelles il est possible de jouer pour l'améliorer.

La deuxième phase est celle de la **conception et de l'analyse *a priori***. Dans cette phase, le chercheur s'appuie sur les analyses préalables ainsi que sur le cadre conceptuel, pour poser des choix concernant les variables identifiées précédemment.

On y décrit les orientations générales que l'on compte suivre ainsi que les choix au niveau local. L'analyse *a priori* est généralement *centrée sur les caractéristiques d'une situation adidactique* que l'on va soumettre aux élèves [Artigue, 1988].

Plus précisément, on peut se poser les questions suivantes :

1. *Quel est le problème que chacun des élèves a en charge de résoudre ?*
2. *Peut-on expliciter ce problème en termes de théorie des jeux ?*
3. *Qu'est-ce qu'il suffit à l'élève de savoir ou savoir faire pour comprendre la consigne (entrer dans le jeu) ?*
4. *Qu'est-ce qu'il suffit à l'élève de savoir ou de savoir faire pour réussir (gagner le jeu) ?*
5. *Quel est le contrôle que l'élève a sur son action ?*
6. *Y a-t-il plusieurs phases ?* [Artigue, 1988]

Cette analyse se focalise généralement autour de la relation élève-situation en laissant de côté le rôle de l'enseignant qui ne sera considéré qu'à travers la dévolution du problème et l'institutionnalisation des connaissances.

La troisième phase, l'**expérimentation**, est celle dans laquelle le dispositif imaginé est testé. Elle peut aussi inclure des données issues d'autres types de méthodologies : questionnaires, entretiens de groupes, *etc.*

Dans la phase d'analyse *a posteriori*, le chercheur décrit les observations qu'il a faites au regard des choix qu'il a posés initialement et interprète les différences entre le prédit et l'observé en termes de validation des hypothèses initiales.

3.1.3 Question de recherche

Formulée de manière très générale, la question qui sous-tend ce chapitre est : comment enseigner l'inférence statistique aux étudiants des filières biomédicales à l'Université de Namur ?

Partant du constat qu'actuellement l'enseignement de l'inférence statistique à ces étudiants repose essentiellement sur le test d'hypothèses et prenant appui sur le cadre conceptuel fourni par la théorie des situations didactiques, nous en arrivons à formuler, de manière plus précise, la question de recherche suivante :

Dans quelle mesure le recours à une situation conçue pour reproduire les conditions d'émergence du test d'hypothèses permet-il aux étudiants des filières biomédicales à l'Université de Namur d'en apprendre la logique sous-jacente ?

En vue de répondre à cette question générale, nous aborderons les sous-questions suivantes :

A quel point la situation mise au point permet-elle d'atteindre les objectifs visés ? Permet-elle de faire émerger des connaissances pouvant servir de base à un apprentissage satisfaisant de la logique sous-tendant le test d'hypothèses ? Si oui, dans quelle mesure ? Et si non, quelles conceptions font obstacle à cet apprentissage ?

Par rapport à ces questions, nous faisons l'hypothèse qu'une situation conçue pour reproduire les conditions d'émergence du test d'hypothèses constitue un point de départ adéquat pour l'apprentissage de la logique sous-jacente au test d'hypothèses.

Pour tenter d'apporter des éléments de réponse à ces questions nous suivrons la méthodologie de l'ingénierie didactique.

Dans les **analyses préalables** nous tenterons d'identifier la logique sous-jacente au test d'hypothèses, nous décrirons l'enseignement usuel et ses effets, nous essayerons d'anticiper les difficultés des étudiants, nous passerons en revue deux expériences similaires et définirons le champ des contraintes agissant sur le dispositif d'enseignement expérimental.

Dans l'étape de **conception**, nous construirons le dispositif d'enseignement en nous basant sur une situation qui se veut a-didactique et fondamentale pour le test d'hypothèse. Les comportements attendus sous l'hypothèse selon laquelle cette situation permet à ces étudiants d'apprendre la logique sous-jacente au test d'hypothèse seront décrit dans l'**analyse *a priori***.

Le dispositif d'enseignement sera mis en œuvre et les observations expérimentales seront décrites dans l'étape d'**expérimentation**.

Les écarts entre les comportements observés et attendus feront l'objet de l'**analyse *a posteriori*** ce qui permettra ensuite d'aborder la question de l'éventuelle validation de notre hypothèse générale.

3.2 Analyses préalables

3.2.1 Analyse épistémologique

Les analyses préalables débutent normalement par l'analyse épistémologique des notions visées par l'enseignement, ici le test d'hypothèses.

Cependant, le chapitre précédent est déjà en grande partie construit autour d'une analyse épistémologique du savoir "savant" relatif aux outils de test statistique dont le test d'hypothèses fait partie. Ainsi nous avons choisi de ne rappeler ici que les éléments nécessaires à l'identification des conditions d'émergence du test d'hypothèses par comparaison avec le test de significativité selon Fisher et un test statistique bayésien basé sur la probabilité *a posteriori*.

Pour ce faire, nous proposons de les comparer au regard des critères suivants.

1. **Les hypothèses** : quelles sont les caractéristiques de l'hypothèse ou des hypothèses en jeu dans le test statistique ?
2. **L'objectif** : quel est l'objectif poursuivi par rapport à cette hypothèse ou à ces hypothèses ?
3. **Le cadre probabiliste** : dans quel cadre probabiliste se situe-t-on ? Considère-t-on que la probabilité est un concept qui peut être utilisé pour mesurer un niveau de croyance dans une hypothèse (cadre bayésien) ou non (cadre fréquentiste) ?
4. **Le moment** : la démarche liée au test statistique intervient-elle avant que les données ne soient récoltées ou après ? En ce sens, la réflexion se fait-elle *a priori* ou *a posteriori* ?

Le **test de significativité** (selon Fisher) est construit autour d'une seule hypothèse statistique avec l'idée de mesurer à quel point les observations corroborent une certaine hypothèse. On pourrait, en ce sens, le décrire comme une extension de la démarche de corroboration-réfutation formalisée par Popper (1934) aux hypothèses statistiques. On peut ajouter à cela l'idée que l'hypothèse statistique doit être formulée de manière précise pour pouvoir être confrontée aux données. Fisher rejette globalement l'idée que le concept de probabilité puisse s'appliquer directement aux hypothèses elles-mêmes et s'inscrit donc à ce titre dans un cadre fréquentiste. Enfin, il semble que cette démarche prenne son sens une fois que l'on dispose à la fois d'une hypothèse statistique précise et d'une série d'observations permettant d'éprouver l'hypothèse. En ce sens, on peut définir que le moment auquel le test de significativité prend son sens est *a posteriori* par rapport à la récolte des observations (voir tableau 3.1).

Le **test d'hypothèses** (selon Neyman et Pearson) met en jeu plusieurs hypothèses (généralement deux, H_0 et H_1) et propose d'utiliser les observations pour faire un choix entre (1) considérer que l'hypothèse H_0 est vraie ou (2) considérer que l'hypothèse H_1 est vraie. Cela implique que parmi les deux hypothèses concurrentes, l'une est correcte. Dans ce cadre, le test d'hypothèses constitue un moyen de contrôler le risque de choisir H_1 lorsque H_0 est vraie (risque d'erreur α) et de minimiser le risque de choisir H_0 lorsqu'en réalité c'est H_1 qui est vraie (risque d'erreur β). A l'instar du test de significativité, le test d'hypothèses s'inscrit dans un cadre probabiliste fréquentiste. Par contre, contrairement au test de significativité, la démarche du test d'hypothèses se déroule essentiellement en amont de la récolte des observations. Il s'agit, par exemple, de réfléchir *a priori* au nombre d'observations qui seraient nécessaires pour garantir un choix correct en s'assurant que les risques d'erreurs α et β restent acceptables (voir tableau 3.1).

Les tests statistiques bayésiens reposent sur une philosophie très différente de celle sous-jacente aux tests statistiques fréquentistes (voir chapitre 2) et en décrire le fonctionnement permet de mieux comprendre ce qui définit l'approche fréquentiste dans laquelle on retrouve

le test d'hypothèses. Nous proposons donc d'inclure à cette comparaison un **test statistique bayésien**. Il existe cependant de nombreux tests statistiques bayésiens [2], nous n'en présentons ici qu'un seul, qui a pour caractéristique de n'être pas trop éloigné de la logique du test d'hypothèses. L'objectif de ce test statistique bayésien serait de mettre à jour le niveau de croyance dans différentes hypothèses concurrentes (par exemple deux hypothèses H_0 et H_1) en se basant sur les observations récoltées. Partant d'un niveau de croyance initiale (la probabilité *a priori* de chacune des hypothèses) et combinant celle-ci aux observations à l'aide du théorème de Bayes, il est possible de déterminer les probabilités *a posteriori* des hypothèses H_0 et H_1. A l'inverse du test de significativité, la réflexion ne se déroule pas entièrement *a posteriori* puisqu'elle nécessite d'avoir défini *a priori* les probabilités associées à chaque hypothèse. Cependant, contrairement au test d'hypothèses, la réflexion ne se fait pas essentiellement *a priori* puisqu'elle est centrée sur une série d'observations particulières et pose la question de savoir comment intégrer ces observations pour mettre à jour le niveau de croyance que l'on peut avoir dans l'une ou l'autre hypothèse. A ce titre, on qualifiera le moment de la réflexion de "mixte", ni essentiellement *a priori*, ni uniquement *a posteriori* (voir tableau 3.1).

Dès lors, pour faire émerger artificiellement le test d'hypothèses il nous apparait nécessaire d'élaborer une situation qui implique la nécessité de **faire un choix entre plusieurs hypothèses statistiques**. Ces hypothèses doivent être précisément définies (il faut connaître les paramètres qui définissent les distributions de probabilités) afin de pouvoir réfléchir aux risques d'erreurs. La question doit être posée **en amont de la récolte des données** et doit, idéalement, **induire une réflexion entre le nombre d'observations à réaliser et le risque d'erreur que cela impliquerait**. Par ailleurs, la situation devrait induire une réflexion dans un cadre fréquentiste ce qui a probablement plus de chances d'arriver si on se situe en amont de la récolte de données et que l'on imagine un contexte qui permettrait **la répétition** à l'envi de la récolte de données plutôt que si l'on se situe en aval de la récolte de données et qu'il convient de tirer un maximum d'informations d'une série d'observations.

Ces conditions seront reprises au moment d'élaborer la situation didactique à soumettre aux étudiants.

Mais avant de passer à la conception de la situation fondamentale, nous allons poursuivre les analyses préalables par l'analyse de l'enseignement usuel et de ses effets.

2. Certains sont basés sur le facteur de Bayes, d'autres sur la probabilité *a posteriori* pour des hypothèses discrètes, d'autres encore décrivent la distribution de probabilité associée au paramètre visé.

TABLE 3.1 – **Comparaison de trois outils de test statistique**

Critère	Test de significativité	Test d'hypothèses	Test statistique bayésien
Hypothèse(s)	Une hypothèse (H)	Deux hypothèses précises (H_0, H_1), dont une sert de référence (H_0)	Deux hypothèses (H_0, H_1) qui peuvent être équiprobables *a priori* ou non
Objectif	Mesurer le niveau auquel les observations corroborent H	Faire un choix entre H_0 et H_1 en contrôlant les risques d'erreur	Mesurer le niveau de croyance à accorder aux hypothèses H_0 et H_1 compte tenu d'un niveau initial et des observations
Cadre probabiliste	Fréquentiste	Fréquentiste	Bayésien
Moment	*A posteriori*	*A priori*	Mixte

3.2.2 Analyse de l'enseignement usuel et de ses effets

L'analyse de l'enseignement usuel, au niveau local, est décrit en détail dans la section 2.4. A nouveau, afin d'éviter les redondances, nous renvoyons le lecteur qui souhaiterait une analyse détaillée de l'enseignement usuel vers cette section et nous n'en présentons ici qu'une vue synthétique.

De cette analyse, il ressort que le test d'hypothèses tel qu'il est actuellement enseigné s'écarte du test d'hypothèses tel qu'initialement décrit par Neyman et Pearson. Si l'on se réfère aux quatre critères qui viennent d'être utilisés pour présenter les trois outils de test statistique, nous pouvons constater que le test d'hypothèses tel qu'enseigné semble constituer un outil hybride entre le test de significativité de Fisher et le test d'hypothèses de Neyman et Pearson (voir tableau 3.2).

En effet, concernant les **hypothèses**, le test d'hypothèses enseigné est construit autour de deux hypothèses dont l'une est définie précisément et correspond à l'absence d'effet (H_0) et dont l'autre correspond à la présence d'effet et n'est pas définie précisément (H_1). De ces deux hypothèses, H_1 constitue généralement l'hypothèse d'intérêt et représente, en quelque sorte, l'équivalent de l'hypothèse H dans le test de significativité.

Du point de vue de l'**objectif** du test d'hypothèses enseigné, il s'agit de voir si les observations réfutent assez H_0 pour pouvoir considérer que H_1 est vraie. Et dans le cas où les observations ne réfutent pas assez H_0 alors on ne peut pas affirmer que H_0 est vraie. Cette manière de procéder s'approche du test de significativité car on y retrouve l'idée que l'on peut réfuter une hypothèse mais difficilement la prouver ou la valider. Elle s'approche également

TABLE 3.2 – **Comparaison entre les caractéristiques des tests statistiques en théorie et celles du test statistique enseigné**

Critère	Test de significativité (selon Fisher)	Test d'hypothèses (selon Neyman et Pearson)	Test d'hypothèses (tel qu'enseigné)
Hypothèse(s)	Une hypothèse (H)	Deux hypothèses précises (H_0, H_1), dont une sert de référence (H_0)	Une hypothèse précise (H_0, absence d'effet) et une hypothèse imprécise (H_1, présence d'effet).
Objectif	Mesurer le niveau auquel les observations corroborent H	Faire un choix entre H_0 et H_1 en contrôlant les risques d'erreur	Déterminer si les données permettent de mettre en évidence un effet en rejetant H_0
Cadre probabiliste	Fréquentiste	Fréquentiste	Fréquentiste
Moment	*A posteriori*	*A priori*	*A posteriori*

du test d'hypothèses selon Neyman et Pearson dans la mesure où l'on retrouve l'idée de faire un choix entre deux conclusions à partir des observations. Dans ce cas-ci, les deux conclusions seront : l'effet est mis en évidence (rejet de H_0) ou l'effet n'est pas mis en évidence (acceptation de H_0).

Quant au moment auquel la réflexion arrive, on se situe plus souvent dans une démarche *a posteriori* par rapport aux observations plutôt qu'*a priori*. En cela, le test d'hypothèses enseigné se rapproche plutôt du test de significativité.

Ainsi, il nous semble que l'enseignement usuel, en tout cas au niveau local, se soit écarté des conditions initiales dans lesquelles le test d'hypothèses était une réponse adaptée. Cela pourrait, peut-être, expliquer les difficultés rencontrées par les étudiants à comprendre la logique du test d'hypothèses. Par conséquent, nous pensons qu'il serait possible d'améliorer l'enseignement du test d'hypothèses en se rapprochant des conditions de validité initiales du test d'hypothèses, à savoir en posant le problème du choix entre deux hypothèses précises en amont de la récolte des observations.

3.2.3 Difficultés attendues

Notre expérience d'enseignement au cours des séances de travaux pratiques nous donne une première idée concernant les points qui posent problème dans l'apprentissage de la démarche du test d'hypothèses. En effet, suivre de près des petits groupes d'étudiants au cours des différentes séances de travaux pratiques permet d'identifier les exercices particulièrement complexes et les

questions qui reviennent de manière systématique.

Nous avons identifié trois sources de difficultés.

La notion de modèle est cruciale en statistique, qu'il s'agisse de décrire une série d'observations ou de faire de l'inférence. En effet, le modèle peut être utilisé, dans l'analyse descriptive, pour résumer une réalité complexe ou bien, dans l'inférence, pour représenter la distribution des résultats attendus sous une certaine hypothèse. Au cours des séances de travaux pratiques, les étudiants passent une grande partie de leur temps à manipuler des modèles statistiques. Les principaux modèles qui sont utilisés en TP sont le modèle binomial, le modèle de Poisson, le modèle normal et le modèle de χ^2.

La première source de difficulté réside dans **le statut du modèle**. Bien que les étudiants passent de nombreuses heures à manipuler des modèles statistiques, ils semblent être nombreux à ne pas bien saisir le rôle de ces outils aussi bien en analyse descriptive qu'en analyse inférentielle. Il est possible que cette incompréhension perdure tout au long des travaux pratiques. En effet, les étudiants peuvent résoudre la plupart des exercices en appliquant des algorithmes, des "recettes de cuisine", sans véritablement saisir ce qui est en jeu à chacune des étapes qu'ils mettent en œuvre.

A côté des difficultés liées au statut du modèle, nous observons beaucoup de difficultés à **manipuler les tables** associées aux modèles théoriques. En effet, celles-ci reprennent les valeurs de la fonction de répartition ($P(X \leq x)$) alors que les étudiants ont plus de facilité à visualiser la fonction de probabilité ($P(X = x)$) (voir figure 3.1 et table 3.3).

TABLE 3.3 – **Fonction de probabilité et fonction de répartition dans le cas d'une distribution discrète**. Fonction de répartition ($P(X = x)$) et fonction de répartition ($P(X \leq x)$) pour une loi binomiale de paramètres $n = 5$ et $\pi = 0.5$.

	$P(X = x)$	$P(X \leq x)$
$x = 0$	0.03125	0.03125
$x = 1$	0.15625	0.18750
$x = 2$	0.31250	0.50000
$x = 3$	0.31250	0.81250
$x = 4$	0.15625	0.96875
$x = 5$	0.03125	1.00000

La troisième difficulté est liée à la manipulation de distributions continues telles que la distribution normale. Deux éléments semblent poser problèmes aux étudiants : le concept de densité de probabilité et celui de distribution d'échantillonnage.

Concernant la densité de probabilité, il faut noter que Calmant (2004) avait, déjà en son

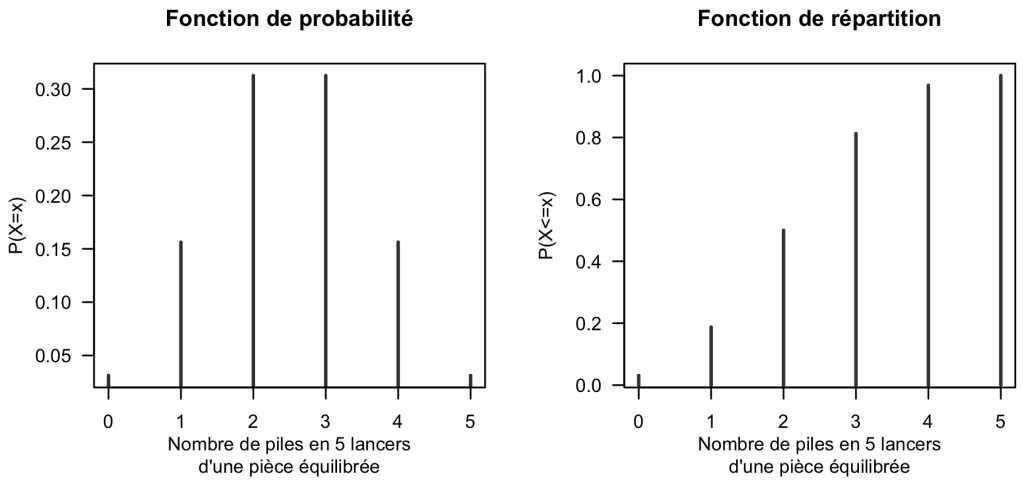


FIGURE 3.1 – **Fonctions de probabilité et de répartition**. Fonction de répartition ($P(X = x)$) et fonction de répartition ($P(X \leq x)$) pour une loi binomiale de paramètres $n = 5$ et $\pi = 0.5$.

temps, décrit les difficultés rencontrées par les étudiants universitaires de différentes sections scientifiques. Il s'était notamment basé sur des entretiens semi-dirigés pour émettre l'hypothèse selon laquelle le dispositif d'enseignement, qui reposait alors en grande partie sur le multimédia, pouvait causer des difficultés d'appréhension de la courbe de Gauss. En effet, ces étudiants étaient très fréquemment confrontés à cette représentation graphique à travers les différentes pages Web du site d'auto-apprentissage mais ne savaient visiblement pas expliquer ce que représentait la courbe lorsque cela leur était demandé [Calmant, 2004].

La difficulté du concept de densité de probabilité réside dans le fait que, lorsque l'on manipule des distributions théoriques continues, telles que la distribution normale, la probabilité d'observer certaines valeurs de la distribution est représentée par la surface sous la courbe. En effet, si on prend le cas de la loi de probabilité qui décrit la distribution, discrète, du nombre de résultats "pile" que l'on peut observer en cinq lancers d'une pièce de monnaie équilibrée, on constate que l'axe des ordonnées est la probabilité d'un certain résultat. Par exemple, la probabilité d'avoir exactement 4 "pile" en 5 lancers est de 16 %. Par contre, si on travaille sur une distribution continue, par exemple une loi normale décrivant la distribution de taille d'individus (en cm), alors la probabilité d'observer un individu mesurant exactement une certaine valeur (180 cm par exemple) tend vers 0 tandis que la probabilité d'observer un individu mesurant entre 180 et 190 cm sera donnée par la surface sous la courbe entre les bornes 180 et 190. Dans le cas d'une distribution continue, l'axe des ordonnées représentera la densité de probabilité et non la probabilité (voir figure 3.2).

Une autre difficulté survenant lorsque l'on manipule les distributions continues est la conceptualisation de la distribution d'échantillonnage. Supposons que l'on travaille sur la distribution

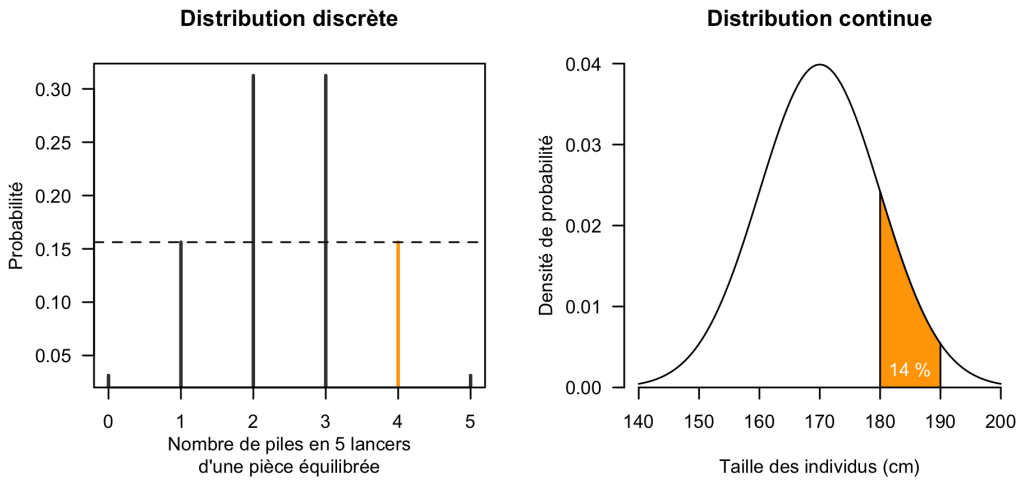


FIGURE 3.2 – **Exemples de distributions discrète et continue**. Gauche : distribution binomiale de paramètre $n = 5$ et $\pi = 0,5$. Droite : Distribution normale de paramètres $\mu = 170cm$ et $\sigma = 10cm$.

normale, à partir de la distribution des valeurs individuelles (trait continu sur la figure 3.3), par exemple de moyenne 170 cm et d'écart-type 10 cm, il est possible de trouver la distribution des moyennes de trois individus tirés aléatoirement dans la population initiale. Cette distribution d'échantillonnage aura une moyenne de 170 cm et un écart-type de $\frac{10cm}{\sqrt{3}} = 5.8$ cm. Si, mathématiquement, les étudiants n'éprouvent pas de difficulté à appliquer la formule permettant de calculer l'écart-type de la distribution d'échantillonnage à partir de celui de la distribution des valeurs individuelles, ils semblent éprouver beaucoup plus de difficultés à comprendre ce que représente la distribution d'échantillonnage. Une distribution d'individus est relativement facile à conceptualiser mais une distribution de moyennes de n individus est plus abstraite et peut être une source de difficulté pour les étudiants.

Parmi les sources de difficultés que nous avons identifiées, certaines sont inhérentes à l'apprentissage de la logique du test d'hypothèses et ne pourraient, selon nous, être évitées. C'est le cas de la notion de modèle statistique.

Par contre, il nous semble possible et souhaitable de concevoir une situation fondamentale pour le test d'hypothèses qui évite les difficultés que constitue la manipulation d'une distribution continue et d'une distribution d'échantillonnage.

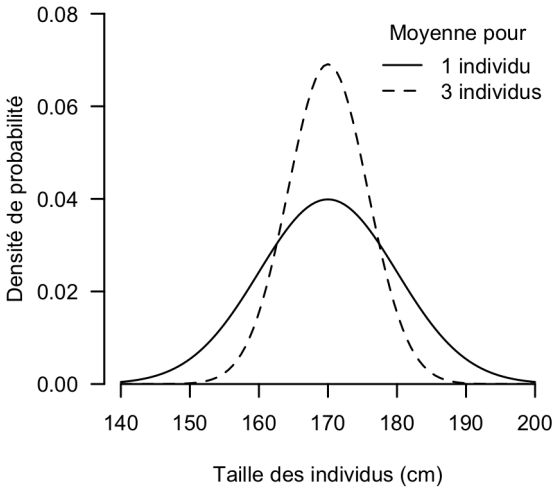


FIGURE 3.3 – **Distribution d'échantillonnage**. Distribution des valeurs individuelles (trait continu) ou des moyennes de trois individus (trait pointillé).

3.2.4 Description d'expériences similaires

Le dispositif expérimental que nous allons mettre au point pour tester notre hypothèse n'est pas le premier à tenter de reproduire les conditions d'émergence de l'inférence statistique ou du test d'hypothèses.

En effet, notre dispositif sera une adaptation du dispositif imaginé et testé par Brousseau (1974) chez des élèves de CM2 (environ 10 ans) et adapté plus tard pour des élèves de 4e au collège (environ 14 ans) par Régnier (1998).

Ainsi, dans ces analyses préalables, il nous paraissait essentiel de revenir sur les expériences similaires et desquelles nous nous sommes inspiré pour construire notre dispositif expérimental.

L'expérience originale (Brousseau, 1974)

De manière intéressante, on peut retrouver chez Brousseau une réflexion sur la situation fondamentale du test d'hypothèses. Il a, en effet, décrit une expérience au cours de laquelle il a testé une situation fondamentale du test d'hypothèses chez des élèves de l'enseignement primaire (décrit dans [Brousseau, 2005]).

Il s'agit d'une situation fondamentale génétique qui aboutit (par la mise en place successive des différentes connaissances) à une situation fondamentale signifiante pour le test d'hypothèses. L'intérêt de cette expérience qui s'est déroulée dans l'enseignement primaire est de démontrer qu'il est possible de placer les élèves face à un problème dont ils se sortiront eux-mêmes en

construisant progressivement un raisonnement qui a toutes les caractéristiques du test d'hypothèses. La situation initale est donc bien une situation génétique car elle permet d'engendrer (avec un minimum d'apports extérieurs) la connaissance d'une part et, d'autre part, la situation qui aboutira à la connaissance suivante. La situation finale est une situation signifiante car la connaissance que vont développer les élèves, bien qu'associée à des méthodes simplifiées, leur permet d'atteindre le sens global du test d'hypothèses.

Le problème soumis initialement aux élèves est le suivant : il s'agit de déterminer le nombre de jetons noirs et blancs que contiennent trois sacs opaques. Les élèves savent que chaque sac contient cinq jetons au total et qu'ils ne peuvent être que noirs ou blancs.

> " *Vous allez essayer maintenant de deviner quelle est la composition de chaque sac. Mais comme on n'a pas le droit de regarder dans les sacs et que personne ne connaît leur contenu exact, personne ne pourra vous dire si vous avez deviné juste. Il faudra vous convaincre vous-même !*" [Brousseau, 2005].

Les élèves peuvent piocher un jeton dans le sac pour observer sa couleur avant de le remettre (tirage avec remise). La particularité du problème est que l'enseignant leur annonce d'emblée qu'il n'ouvrira jamais les sacs, ce qui met les élèves dans une situation proche de celle du chercheur qui doit affirmer quelque chose sur le monde dans un contexte où il n'existe pas de certitudes. Il peut multiplier les prises d'informations (ce qui a un certain coût) mais jamais arriver à une certitude absolue. Ainsi, l'enjeu pour les élèves n'est pas tant de savoir si le sac contient zéro, un, deux, trois, quatre ou cinq jetons blancs mais plutôt d'imaginer et de mettre à l'épreuve un raisonnement qui leur permette d'arriver à tirer des conclusions à propos d'un objet inconnu duquel ils ne peuvent avoir que des informations parcellaires. On voit bien ici qu'il s'agit d'une situation fondamentale signifiante pour le test d'hypothèses.

Très vite, les élèves émettent des hypothèses que l'enseignant invite à vérifier expérimentalement, ce qui permet, au passage, de mettre à mal certaines conceptions erronées à propos des probabilités [3]. Ils se disent ensuite qu'ils doivent noter les résultats. Ceux-ci sont enregistrés d'abord sous forme de fréquences des séries de cinq jetons, les élèves compteront plus tard le rapport entre le nombre de jetons blancs et le nombre de jetons observées. Ils découvriront ainsi, sans le savoir, les caractéristiques, fluctuantes, d'une statistique.

Par la suite, les élèves ont l'idée de construire eux mêmes des machines de hasard. L'enseignant leur fournira pour cela des bouteilles, transparentes cette fois-ci, qui contiennent un nombre de billes bleues ou jaunes déterminé afin d'analyser les tirages que l'on obtiendrait dans les différentes situations existantes. Sans le savoir, les élèves ont construit des modèles qu'il maitrisent afin de connaitre la distribution des résultats qu'ils pourraient obtenir sous chacune de ces hypothèses, ce qui est exactement la démarche utilisée dans le test d'hypothèses. Avec l'aide

3. L'existence d'une loi de *compensation* par exemple.

de simulations informatiques, les élèves ont la possibilité d'étudier le comportement de ces modèles et, en particulier, le fait que la proportion de billes bleues est de moins en moins variable à mesure que le nombre d'observation est important. Autrement dit, plus le nombre de tirages augmente, plus cette proportion semble se stabiliser autour d'une certaine valeur (qui est la probabilité au sens fréquentiste).

Lors des 31 séances qui forment cette expérimentation, les élèves construiront progressivement un raisonnement statistique complexe. Au cours de cette lente progression, on peut épingler les six jalons suivants dans le raisonnement (librement adapté de [Brousseau, 2005]).

1. **Déterminisme**. Parmi les six contenus possibles, les élèves en évacuent rapidement deux à l'aide d'un raisonnement déterministe simple : il y a au moins un jeton de chaque couleur dans chacun des sacs.

2. **Effectifs**. Pour avoir une idée du contenu du sac, chaque élève réalise un tirage. Les effectifs obtenus à partir de chacun des sacs sont comparés : 9N8B(9 noirs et 8 blancs) pour le sac A , 11B6N pour le sac B et 5N12B pour le sac C [4].

3. **Variabilité**. Les tirages sont recommencés le lendemain et les résultats sont comparés à ceux de la veille. Cette fois, c'est 10N7B pour A, 12N5B pour B et 4N13B pour C. D'un tirage à l'autre, la composition varie.

   "*E4 : Hier, il y avait 9 noirs pour A, là, il y en a 10, dans tous les sacs il y a 1 de différence.* (*ibid*).

   En regroupant les tirages par paquets de 5, les élèves observent tantôt 0B5N, 1B4N, 2B3N, 3B2N, 4B1N ou 5B0N. Pour un même sac, ils mesurent à quel point les compositions observées peuvent varier et observent que certaines compositions reviennent plus souvent que d'autres (3B2N et 4B1N reviennent plus souvent). Notons qu'ils éliminent les observations 5B0N et 5N0B, jugeant que ces compositions sont, de toute façon, impossibles vu que les deux couleurs ont été observées. Ces observations sont, dans un premier temps, considérées comme des erreurs et les tirages sont recommencés.
   Ce stade de la réflexion est observé dès la deuxième séance.

4. **Modèle**. A la sixième séance arrive l'idée de construire un modèle, d'utiliser une machine de hasard dont on connaitrait la composition et dont on pourrait étudier le comportement.

   "*Un enfant propose alors de fabriquer un sac dont on connaîtrait la composition. On leur présente un nouveau matériel : une bouteille et des billes* [5]*. On*

4. ce qui représente un tirage par élève dans chacun des sacs.
5. Dispositif jugé plus pratique pour effectuer de nombreux tirages.

met dans la bouteille 4 billes bleues et une jaune. On l'appelle la bouteille Z. [Brousseau et al., 1974].

Les élèves utilisent la bouteille Z pour observer la distribution des résultats qui serait obtenue sous une hypothèse précise. Notons qu'à partir de ce moment, pour des raisons pratiques, les sacs A, B et C seront remplacés par des bouteilles opaques et des billes. Ils commencent à réaliser un grand nombre de tirages dans chacune des bouteilles A, B et C et notent les résultats dans un tableau (voir tableau 3.4).

TABLE 3.4 – **Représentation des résultats des tirages dans la bouteille A**. Le tableau reprend les fréquences de billes noires et blanches ainsi que les fréquences relatives cumulées. Les étoiles indiquent les calculs erronés. Source : [Brousseau, 2005]

série	1	2	3	4	5	6	7	8	9	10	11	12
Blancs	5	2	7	4	3	4	3	3	3	6	5	5
Cumul	5	7	14	18	21	25	28	31	34	40	45	50
Noirs	5	3	8	6	7	5	7	7	7	4	5	5
Cumul	5	8	16	22	29	34	41	48	55	59	64	69
Nbr.tir	10	15	30	40	50	59	69	79	89	99	109	119
.												
Calcul	5 :10	7 :15	14:30	18:40	21:50	25:59	28:69	31:79	34:89	40:99	45:109	50:119
Fréq.B	0,5	0,45	0,49	0,45	0,42	0,42	0,40	0,39	0,48*	0,44*	0,41	0,46*
Calcul	5 :10	8 :15	16 :30	22 :40	29 :50	34 :59	41 :69	48 :79	55 :89	59 :99	64:109	69:119
Fréq.N	0,50	0,53	0,53	0,55	0,58	0,57	0,57*	0,60	0,61	0,50*	0,58	0,57

Plus loin (entre les 13e et 15e séances), les élèves font de grandes séries de tirages sur des bouteilles modèles comprenant 1B4N, 2B3N, 3B2N et 4B1N. A nouveau, ils calculent la fréquence cumulée. Cette fois-ci, ils sont invités à reporter l'évolution de cette fréquence cumulée sur graphique (voir figure 3.4). Ils découvrent ainsi la probabilité (au sens fréquentiste) comme la valeur théorique vers laquelle tend la fréquence relative lorsque le nombre de tirage tend vers l'infini.

A la 16e séance, un ordinateur muni d'un programme permettant de simuler de grandes séries de tirages à partir des différentes compositions possibles est mis à disposition des élèves. Il permettra d'accélérer le processus de tirage qui est devenu long et fastidieux pour les enfants. Sur base de ces modèles, les élèves constatent qu'à partir d'un certain nombre d'observations, le rapport du nombre de boules blanches sur le nombre de boules total se stabilise (il semble entrer dans un intervalle, une bande et y rester).

5. **Intervalles de décision**. Par la suite, la maîtresse introduit un nouveau jeu (jeu des devinettes). Les élèves disposent d'un certain crédit en jetons, ils peuvent les dépenser pour réaliser des tirages puis doivent deviner le contenu d'une bouteille préparée par la maîtresse. Plus ils réalisent de tirages, plus ils ont de chances de gagner mais moins le gain est important. L'information a un coût. A l'issue de ce jeu, elle demande aux élèves

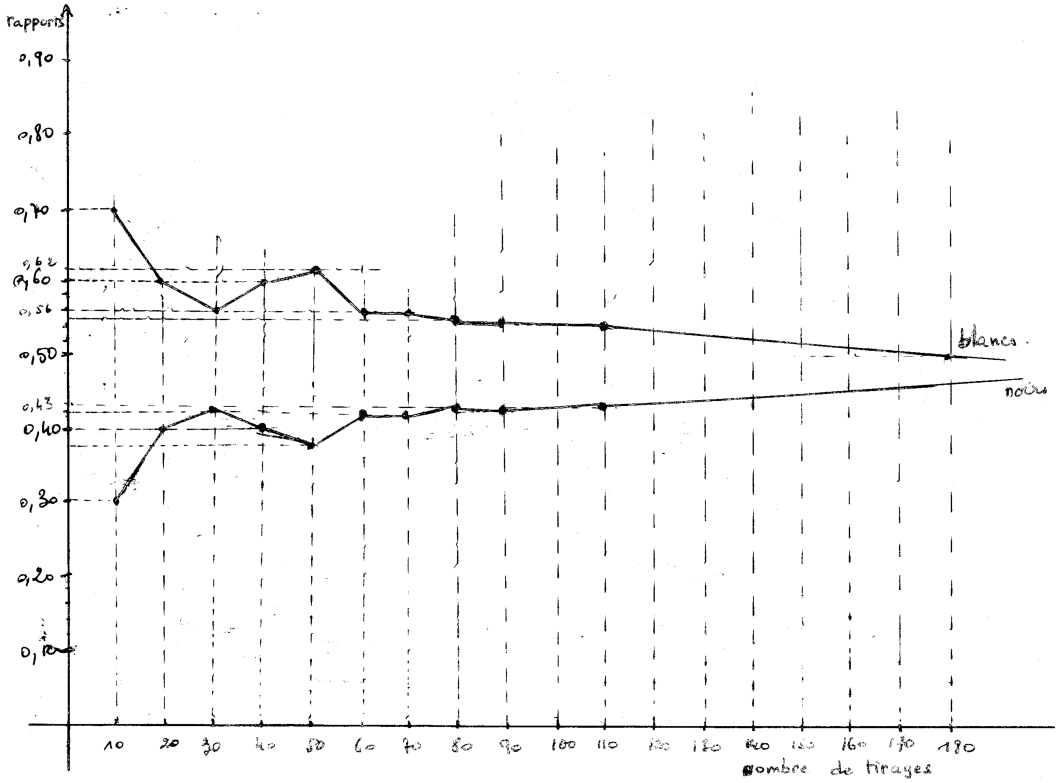


FIGURE 3.4 – Graphique représentant l'évolution du rapport entre les boules de différentes couleurs et le total des observations en fonction du nombre de tirages simulés par ordinateur. Schéma issu de [Brousseau, 2005].

TABLE 3.5 – **Intervalles de décision**

Proportion de blancs	Composition considérée correcte
0	0B5N
entre 0,01 et 0,29	1B4N
entre 0,30 et 0,49	2B3N
entre 0,50 et 0,69	3B2N
entre 0,70 et 0,99	4B1N
1	5B0N

comment ceux-ci s'y prennent pour décider quelle composition est la bonne à partir des tirages et s'ils peuvent communiquer cette règle à un élève qui n'aurait pas suivi tout le développement.

A travers cette situation de formulation, les élèves sont amenés à identifier des intervalles liés à des décisions. Après quelques débats, les élèves se mettent d'accord sur les intervalles suivants (voir tableau 3.5).

6. **Risque d'erreur et taille des séries**. Au cours de la 24e séance, les élèves repartent des intervalles définis précédemment et ont pour consigne d'*éprouver* ces intervalles, c'est-à-dire de voir, en fonction du nombre de tirages, la proportion de fois où ceux-ci sont corrects. Pour cela, ils utilisent les simulations fournies par l'ordinateur, et vérifient pour plusieurs séries de 10, 20, 50, 200 et 5000 tirages.

   En demandant 15 séries de 15 tirages, 8 séries seulement sont dans la bande et 7 à l'extérieur, avec 15 séries de 20, on réussit 9 fois et on échoue 6 fois, avec des séries de 100, 15 réussites sur 15 mais avec 160 seulement 14 réussites sur 15. Les séries de 1000 tirages se resserrent beaucoup autour de la valeur théorique... Au vocabulaire près, les élèves utilisent la notion de seuil de signification et d'intervalle de décision [Brousseau, 2005].

   Le reste des séances (de la 26e à la 31e séance) est consacré à des expériences plus classiques sur les évènements et les probabilités.

Adaptation pour élèves du lycée (Régnier, 1998)

La situation fondamentale de Brousseau a été adaptée pour des élèves de 4e (13-14 ans), au collège (Régnier, 1998).

Cette adaptation visait à faire émerger des notions telles que le *sondage*, l'*échantillon*, la *statistique*, le *risque*, l'*estimation* ou encore le *test d'hypothèses*.

TABLE 3.6 – **Six jalons dans le raisonnement des élèves**.

	Jalon
1.	Déterminisme
2.	Effectifs
3.	Variabilité
4.	Modèle
5.	Intervalle de décision
6.	Risque d'erreur et taille des séries

La situation était proposée à un groupe de 24 élèves répartis en 4 groupes de 6. Chaque groupe disposait d'une urne composée de 200 billes dont une certaine proportion de billes bleues, inconnue. L'objectif général était de réussir à estimer le plus précisément possible le contenu de l'urne tout en réalisant le minimum de prélèvement.

Selon les auteurs, des indices de l'apprentissage des notions en question peuvent être relevés dans la réaction des élèves au critère de réussite.

> "*Les élèves auront intériorisé quelque chose du raisonnement statistique inférentiel si des protestations surgissent contre l'objectivité du principe pour établir le groupe victorieux à savoir donner un résultat aussi proche que possible de la vraie valeur qui est ici accessible par comptable direct, en ayant dépensé le moins possible. En effet, selon la théorie des sondages et celle des échantillonnages d'une proportion, une démarche rigoureuse et logiquement exacte peut conduire à un résultat faux obtenu à partir de la valeur empirique de la proportion calculée sur un échantillon représentatif, au sens d'obtenu aléatoirement avec ou sans remise*" [Régnier, 1998].

La séquence didactique était décomposée en quatre phases, le tout prévu pour durer 2h :

1. Phase 1 : Chaque élève tire un échantillon de la taille de son choix pour faire une estimation du contenu de l'urne.
2. Phase 2a : Chaque groupe doit se mettre d'accord pour estimer le contenu de son urne.
3. Phase 2b : Chaque groupe doit prendre une décision par rapport à une affirmation concernant le contenu de l'urne.
4. Phase 3 : Chaque groupe doit prendre une décision par rapport à une affirmation concernant l'égalité de contenu entre deux urnes.

L'analyse des raisonnements mis en œuvre révèle que les élèves se basent encore souvent sur l'intuition ou sur la comparaison directe des proportions observées (par exemple, dans deux

urnes différentes) pour prendre leur décision (par exemple, quant à l'égalité ou non du contenu des deux urnes).

Durant la séquence, les élèves ont également à répondre à un questionnaire concernant les termes *sondage*, *échantillon* et *statistique*. Celui-ci montre que les élèves sont confrontés à ces termes et ce dans différents contextes. Le terme *sondage* semble couramment rencontré à la télévision et au collège, le terme *échantillon* évoque plutôt les magazines et les magasins tandis que le terme *statistique* renvoie à des livres scolaires, à la télévision ou encore au collège. Cette analyse soulève les contextes dans lesquels il faudra chercher les conceptions préalables des élèves au moment d'enseigner les notions de sondage, d'échantillon ou de statistique.

Comparaison des caractéristiques des deux expériences

Les deux dispositifs expérimentaux décrits précédemment sont fort intéressants et il nous semble important d'analyser les différences entre les deux dispositifs afin de pouvoir mettre au point un dispositif le plus à même de mettre notre hypothèse à l'épreuve.

Les deux dispositifs diffèrent à plusieurs égards (voir tableau 3.7).

TABLE 3.7 – **Comparaison des caractéristiques des deux dispositifs expérimentaux**. La situation 1 renvoie au dispositif imaginé par Brousseau, la situation 2 renvoie au dispositif expérimenté par Régnier.

Critère	**Situation** 1	**Situation** 2
Groupe	17 élèves	4 x 6 élèves
Age du public	10 ans	13 ans
Durée de la séquence	31 x environ 10 minutes	2 heures
Tirages	Avec remise	Avec ou sans remise
Contenu	5 objets	200 objets
Vérification	Jamais	A la fin de la séquence

Tout d'abord la taille du groupe, l'âge des élèves et la durée de la séquence qui diffèrent. Dans le dispositif que nous allons tester, ces éléments sont pris comme des contraintes extérieures avec lesquelles il faut composer.

Ensuite, le type de tirage est différent. Dans l'expérience de Brousseau, il s'agit d'un tirage avec remise tandis que dans l'adaptation proposée par Régnier, il peut-être avec ou sans remise, au choix de l'élève.

Deux autres différences sont, selon nous, de nature à influencer fortement le type de démarche mise en œuvre.

D'une part il s'agit du nombre d'objets contenus dans le contenant. Avec 5 objets, le nombre d'hypothèses possibles est réduit (six pour être précis) et la question posée est plus naturellement une question de choix entre plusieurs hypothèses. Avec un grand nombre d'objets (200 billes dans chaque urne), le nombre d'hypothèses possibles est lui aussi très grand et la question posée est plus naturellement une question d'estimation (quelles sont les hypothèses compatibles avec les données ?) plutôt qu'une question de test d'hypothèses (quelle hypothèse faut-il considérer correcte au vu des données ?). Ceci dit, il est possible, avec un grand nombre d'objets, d'orienter la question de manière à ce que le test d'hypothèses soit la réponse la plus adaptée (voir phase 2b et phase 3 du dispositif de Régnier).

D'autre part le fait de vérifier ou non le contenu réel de la bouteille ou de l'urne après que les élèves ont donné leurs estimations. Dans le dispositif de Régnier, les estimations sont comparées au contenu réel des urnes en fin de séquence. Dans l'expérimentation de Brousseau, une des caractéristiques essentielles est que ce contenu initial n'est jamais révélé [6]. La question n'est donc pas de savoir si on est proche ou non du contenu réel, mais plutôt de mettre au point et de justifier une manière de faire des prélèvement qui permette de déterminer le contenu exact de la bouteille sans se tromper trop souvent.

3.2.5 Champ de contraintes

Pour construire un dispositif expérimental d'enseignement, il est nécessaire d'identifier le champ des contraintes auquel ce dispositif est soumis. Celles-ci concernent le public, le dispositif d'enseignement dans le cadre duquel l'expérimentation sera réalisée ou encore la chronologie entre le cours et le dispositif expérimental.

En ce qui concerne le public, nous allons nous intéresser aux étudiants universitaires suivant un bachelier en sciences biomédicales, pharmacie et médecine. Le cours de biostatistique intervient, en principe, lors de la deuxième année pour les étudiants en sciences biomédicales et pharmacie et en première année pour les étudiants en médecine. Il faut noter que la plupart de ces étudiants ne se sentent pas particulièrement à l'aise avec les outils mathématiques et que le choix de la filière est parfois motivé précisément par le fait que celle-ci ne repose pas trop sur les mathématiques.

Pour ce public, le raisonnement mathématique est peu développé au cours de la formation puisque les étudiants en sciences biomédicales auront, sur leur 180 crédits de bachelier, un cours de mathématiques (3 ECTS [7]), un cours de biostatistique (4 ECTS) et un cours d'épidémiologie médicale (2 ECTS). Les étudiants en pharmacie et en médecine [8] reçoivent les mêmes cours de

6. On trouvera toutefois, dans le "jeu des devinettes", une comparaison entre les prédictions et le contenu de certaines bouteilles, mais pour l'activité initiale, ce n'est pas le cas.

7. European Credit Transfer and Accumulation System

8. Notons que les étudiants en médecine ont passé avec succès un examen d'entrée portant notamment sur

biostatistique et d'épidémiologie médicale mais ne suivent pas le cours de mathématiques.

Ces étudiants auront donc entre un vingtième (en sciences biomédicales) et un trentième (pharmacie et médecine) de leur formation dévolu à des cours liés de près ou de loin aux mathématiques et à la statistique. Malgré les faibles moyens consacrés à ce volet de leur formation, les attentes sont relativement élevées puisque les étudiants sont censés, à l'issue de leur bachelier, être capables de comprendre les notions statistiques présentes dans les articles scientifiques.

Cet écart entre les attentes qui reposent sur la formation en statistique (au sens large) et les moyens qui lui sont alloués représente certainement une des causes des difficultés d'enseignement auxquelles nous sommes confrontés aujourd'hui. Toutefois, dans le présent travail, nous considèrerons les attentes et les moyens comme des contraintes qu'il faut prendre en compte pour élaborer un dispositif expérimental.

A côtés des contraintes liées au public concerné, il existe des contraintes liées au cours de statistique dans le cadre duquel ces expérimentations auront lieu. La logique du cours est présentée plus en détail dans la section 2.4, l'idée ici est de tenter d'identifier les éléments qui vont influencer le dispositif expérimental.

En premier lieu, le choix de la matière à enseigner, avec notamment le test d'hypothèses comme outil d'inférence statistique est une contrainte majeure puisqu'elle exclut le test de significativité ainsi que les approches bayésiennes.

Une autre contrainte est le moment auquel l'expérience va se dérouler. Elle aura lieu lors des séances de travaux pratiques dont le format est déterminé à l'avance, à savoir, quatre séances de deux heures par groupes d'une vingtaine d'étudiants. Ces séances de travaux pratiques ont pour objectif d'introduire les notions qui seront vues au cours ou bien de les illustrer si elles ont lieu avant celui-ci. L'idée est de balayer l'étendue de la matière au cours de ces séances, c'est-à-dire qu'il faut y aborder l'analyse descriptive jusqu'aux tests d'hypothèses.

Il faut également noter qu'il existe des contraintes au niveau de l'équipe d'assistants qui assurent les séances de travaux pratiques. En effet, il n'est pas toujours facile de modifier les habitudes d'enseignement d'une équipe qui travaille généralement selon une certaine logique (approche que l'on pourrait qualifier de plus classique, non basée sur les principes de la théories des situations didactiques).

Enfin, parmi les contraintes, il faut aussi souligner le type d'évaluation lié à ce cours. Pour les étudiants en sciences biomédicales et en pharmacie, le système d'évaluation est double. Il consiste, d'une part, en une évaluation continue dispensatoire pour l'examen de fin de quadrimestre : trois interrogations par QCM sont réparties sur le quadrimestre de cours et portent sur la matière vue jusqu'alors. D'autre part, un examen de fin de quadrimestre se déroule durant la session d'examen sous la forme d'un questionnaire QCM. Les étudiants en médecine ne

des notions mathématiques vues dans l'enseignement secondaire.

disposent, pour leur part, que de l'examen de fin de quadrimestre.

Quelle que soit la filière, l'évaluation se déroule par QCM, ce qui implique que les étudiants doivent pouvoir trouver la bonne réponse parmi une liste de plusieurs propositions mais ne doivent pas justifier la réponse qu'ils ont choisie. Telle qu'elle est pratiquée dans le cadre de ce cours de biostatistique, l'évaluation par QCM implique également qu'il n'y a pas de questions totalement inattendue ou ayant un format nouveau lors de l'examen. En effet, chaque type de question posé à l'évaluation certificative se retrouve au moins une fois dans les questions proposées aux étudiants pour l'évaluation formative.

3.3 Conception et analyse *a priori*

Dans cette section-ci, nous allons tenter de montrer, dans un premier temps, de quoi le dispositif est composé et ce qui a motivé ce type de dispositif plutôt qu'un autre. Dans un second temps, dans l'analyse *a priori*, nous essayerons de décrire ce qui, dans le dispositif est censé être de nature à provoquer, chez les étudiants, l'émergence des connaissances visées.

Le dispositif expérimental consiste en une série de quatre séances de travaux pratiques de deux heures chacune. La séance qui nous intéresse particulièrement est la troisième dans laquelle les étudiants seront confrontés à une situation fondamentale pour le test d'hypothèses. Cependant, il nous semble important de passer en revue le contenu et les méthodes appliquées aux séances 1 et 2 car elles contribuent à installer le contrat didactique qui sera d'application lors de la troisième séance et elles ont pour objectif d'aider les étudiants à surmonter certaines difficultés attendues dans la situation fondamentale du test d'hypothèses. La quatrième séance aura pour objectif d'institutionnaliser les connaissances construites durant la troisième séance.

3.3.1 Séance 1 : L'analyse descriptive

La première séance de travaux pratiques ne concerne pas l'inférence statistique mais, au vu des contraintes énoncées plus haut, nous ne pouvions faire l'impasse sur une séance abordant l'analyse descriptive car un des objectifs des séances de travaux pratiques est d'aborder les principaux points de matière.

La situation proposée aux étudiants est conçue pour faire émerger ou mobiliser certains concepts permettant de mesurer la tendance centrale (la médiane) et variabilité (l'écart-type) au sein d'une série de données ainsi que pour favoriser la représentation graphique d'une distribution de valeurs à l'aide d'un histogramme. Pour la majorité des étudiants [9], ces concepts n'ont pas encore été abordés au cours théorique.

9. Les étudiants n'ont pas tous leur séance de travaux pratiques en même temps.

Pour ce faire, les étudiants avaient pour consigne de comparer trois séries de valeurs (artificielles) représentant chacune l'évolution du poids de 1000 personnes soumises à un certain régime (A, B ou C). Cependant, les variables didactiques sont telles que les concepts intuitifs pour les étudiants (la moyenne, le minimum ou le maximum d'une série) ne permettent pas de faire la différence entre les trois séries de données (voir figure 3.5). Par ailleurs, le grand nombre de valeurs rend impossible la comparaison des trois séries sans passer par le calcul de valeurs résumées ou sans passer par un graphique.

Après avoir engagé leurs conceptions dans la résolution du problème, les étudiants sont amenés à considérer une autre mesure de la tendance centrale (la médiane) et d'autres mesures de dispersion (l'écart-type). Ces mesures nouvelles (ou pas encore tout à fait maitrisées) viennent éclairer la comparaison entre les trois séries de valeurs : si les trois régimes affichent une même perte de poids moyenne, le régime B semble beaucoup plus variable que le régime A (voir figure 3.5).

Ensuite, l'utilisation d'un graphique représentant la distribution des valeurs permet de comprendre définitivement ce qui différencie les trois régimes (voir figure 3.5).

La consigne est formulée en termes compréhensibles pour les étudiants de sorte qu'ils sont, en principe, en mesure de juger de la qualité du concept utilisé pour répondre à la question posée. Au cours de la séance, l'enseignant est donc censé assurer la dévolution du problème aux étudiants. A deux reprises, l'enseignant devra institutionnaliser les connaissances nouvelles à l'aide d'une mise en commun des approches des différents groupes d'étudiants et de la formalisation des notions utilisées.

Regime	Moyenne	Min	Max	Médiane	Ecart-type
A	−3.87	−19.1	8.36	−3.88	4.02
B	−3.87	−19.1	8.36	−4.30	7.00
C	−3.87	−19.1	8.36	−6.88	6.40

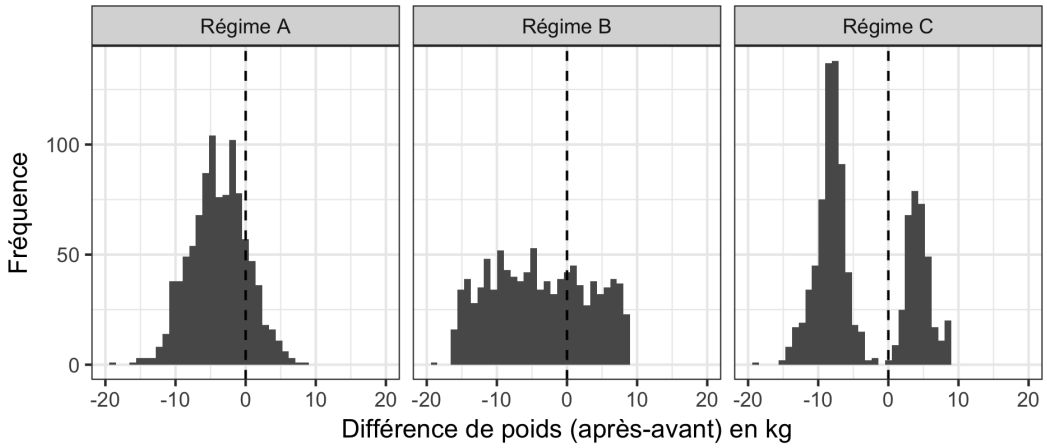


FIGURE 3.5 – **TP1 : tableau descriptif et histogramme pour chacun des groupes comparés**. **Haut** : Tableau descriptif comportant des paramètres familiers pour les étudiants (moyenne, minimum et maximum) qui ne permettent pas de saisir les différences entre les trois jeux de données et des paramètres moins familiers (médiane et écart-type) pour ces étudiants qui, eux, permettent de différencier les trois séries de données. **Bas** : Histogramme pour chaque série de données.

3.3.2 Séance 2 : Modèle binomial

Dans cette séance-ci, nous avons opté pour une approche classique, non basée sur les principes de la théorie des situations didactiques, en réutilisant une séance déjà donnée les années précédentes. Ce choix est basé sur les contraintes pédagogiques car il nous semblait important de ne pas modifier l'ensemble du dispositif d'enseignement pour l'équipe d'assistants. Ne pas modifier cette séance-ci permet aux assistants d'avoir des travaux pratiques plus structurés, plus faciles à contrôler et dont ils ont déjà une certaine expérience.

Par conséquent il ne nous est pas possible d'exposer les objectifs et les comportements attendus des étudiants en utilisant les principes de la théorie des situations didactiques. Nous nous contenterons donc de présenter les objectifs et comportements attendus en termes assez généraux.

Les objectifs sont :

1. Formaliser une série de notions (distribution empirique, distribution théorique, simulations, arbre de probabilité, fonction de répartition et fonction de probabilité) ;
2. Travailler l'utilisation des formules et des tables binomiales ;

3. Entamer une réflexion sur les rapports entre les distributions empiriques et les distributions théoriques.

Le déroulement de cette séance est le suivant. Dans un premier temps, les étudiants récoltent des observations empiriques (lancers de pièces de monnaie) dans le but de construire une distribution expérimentale d'une variable aléatoire. Cette distribution est ensuite comparée à la distribution binomiale (voir figure 3.6). Des simulations sur tableur montrent aux étudiants que les écarts entre la distribution empirique et la distribution théorique se réduisent à mesure que la taille d'échantillon augmente. Dans un deuxième temps, l'assistant montre aux étudiants comment manipuler les formules et tables binomiales et ceux-ci s'exercent en résolvant de courts exercices. Dans un troisième temps, les étudiants travaillent sur la distribution de Poisson, comparant à nouveau la distribution expérimentale (dénombrement de cellules sur un frottis sanguin) à la distribution théorique. Une série de questions récapitulatives a pour objectif d'aider à la formalisation des notions manipulées.

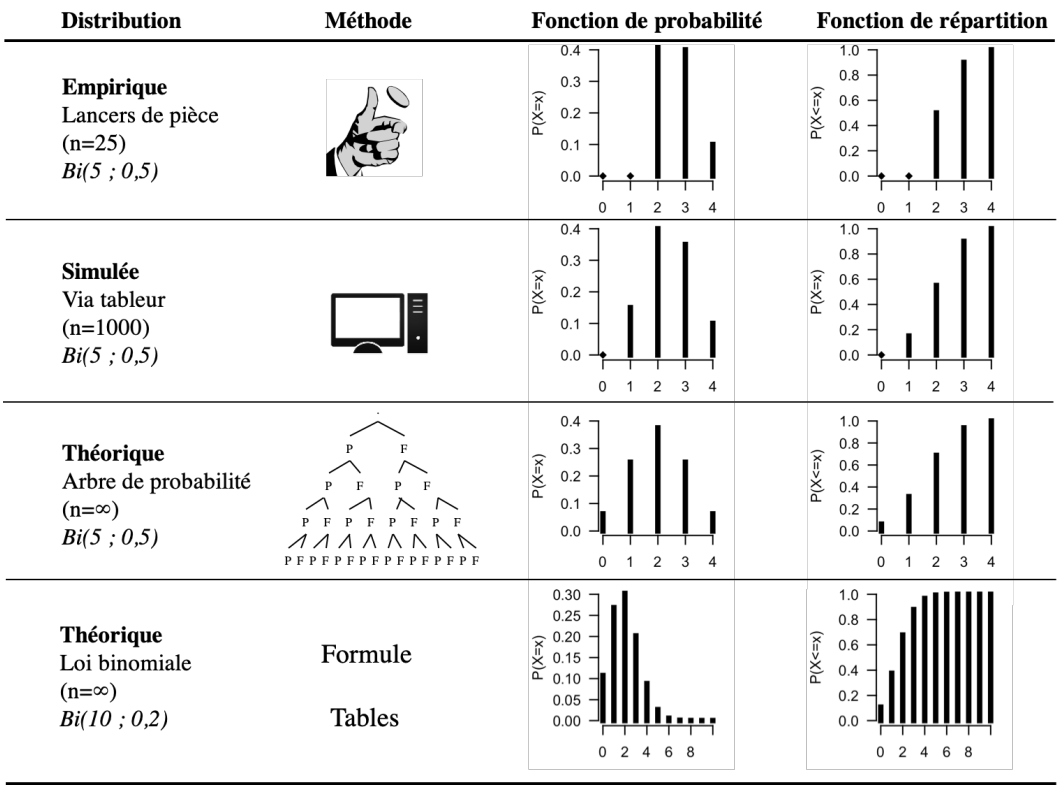


FIGURE 3.6 – **Schéma général de la deuxième séance**.

3.3.3 Séance 3 : Le test d'hypothèses

La troisième séance de travaux pratiques aborde la logique du test d'hypothèses. Cette séance est programmée avant que la matière n'ait été abordée au cours théorique et il n'est pas demandé aux étudiants de l'avoir découverte par eux-mêmes.

Dans ce qui suit, nous allons aborder les caractéristiques de la situation qui a été utilisée, les consignes données aux étudiants ainsi que les raisonnements possibles.

Caractéristiques de la situation

Comme décrit précédemment, pour favoriser l'émergence d'un raisonnement proche de celui sous-jacent au test d'hypothèse, la situation didactique devrait réunir les caractéristiques suivantes :

1. **Il faut faire un choix** : l'étudiant est placé dans une position où les données doivent être utilisées pour faire un choix entre plusieurs hypothèses ;
2. ***...a priori*** : la réflexion se situe en amont de l'expérience, avant que les données ne soient récoltées ;
3. **... basé sur la notion de puissance** : ce choix doit être basé sur l'étude de la relation entre un type de test statistique ou une taille d'échantillon et le risque d'erreur β ;
4. **... et dans un cadre fréquentiste** : et ce, dans un cadre probabiliste fréquentiste.

Au vu de ces caractéristiques, la situation fondamentale proposée par Brousseau à des enfants d'une dizaine d'années, parfois appelée la "bouteille de Brousseau (voir section 3.2.4), nous semble être un bon point de départ.

Cette situation peut être considérée comme une situation de choix puisqu'elle revient à demander aux enfants de trancher, entre six hypothèses [10] à partir des observations. Ou, plus précisément, cela revient à leur demander de mettre au point une méthode qui permettrait de faire ce choix. Il ne suffit pas de faire un pari et de voir si on a visé juste, il faut construire une méthode et la justifier. Le fait qu'il n'y ait qu'un petit nombre d'hypothèses possibles amène à voir la situation comme un problème de décision entre ces différentes hypothèses. Si, par contre, il y avait eu 100 boules parmi lesquelles entre 0 et 100 boules rouges, le nombre d'hypothèses en jeu aurait été trop important et le problème aurait été mieux abordé à l'aide d'outils d'estimation tels que les intervalles de confiance, par exemple.

La situation n'est pas une réflexion *a posteriori* sur base de résultats immuables. Elle est plutôt une réflexion *a priori* sur la méthode à mettre au point pour utiliser les résultats futurs.

La situation pousse également les apprenants à réfléchir au risque d'erreur. En effet, dans l'expérience de Brousseau, les enfants en arrivent à construire leurs propres modèles théoriques, à définir des intervalles équivalents à des zones d'acceptation d'hypothèses et à décrire, sur ces modèles, les proportions de résultats qui tomberaient en dehors de ces zones d'acceptation en prenant des séries de 15, 20, 1000 tirages [Brousseau, 2005].

10. La bouteille contient 0, 1, 2, 3, 4 ou 5 boules rouges.

Enfin, cette situation entre typiquement dans un cadre fréquentiste. La structure de la bouteille et le fait de pouvoir répéter les tirages à l'envi implique presque nécessairement une définition de la probabilité comme la limite d'une fréquence relative quand le nombre de tirages tend vers l'infini. La situation se prête donc assez peu à une définition de la probabilité comme un degré de croyance dans l'une ou l'autre des hypothèses. Notons qu'il aurait été possible d'orienter le problème pour rendre efficace une définition bayésienne de la probabilité.

Brousseau a montré que, chez des enfants de CM2, il était possible au départ de cette situation de faire émerger des notions d'inférence statistique complexes finalement assez proches de celles que l'on rencontre dans le test d'hypothèses. Au vu des caractéristiques du public universitaire auquel on s'adresse, c'est-à-dire des étudiants des filières biomédicales pauvres en mathématiques, il nous a semblé important de travailler l'inférence statistique au départ d'une situation didactique relativement accessible, dans le sens où elle ne fait pas intervenir de notions mathématiques trop complexes.

Par ailleurs, la situation proposée par Brousseau est également intéressant en ce qu'elle évite une série de difficultés que nous avions anticipées et qui sont liées aux distributions continues (voir plus haut), notamment les concepts de densité de probabilité et de distribution d'échantillonnage. En effet, la bouteille de Brousseau est une situation dans laquelle la distribution binomiale est particulièrement adaptée.

Consignes

Nous avons vu que la situation fondamentale imaginée par Brousseau correspondait bien aux caractéristiques requises pour la situation à proposer aux étudiants. Nous avons donc décidé de partir des mêmes consignes que lui.

On présente aux étudiants un contenant opaque contenant cinq objets qui peuvent être de deux types différents [11]. Il est possible de réaliser un tirage avec remise d'un objet à la fois. Le but est de déterminer le contenu de ce contenant sans jamais l'ouvrir.

A partir de là, les étudiants sont invités à raisonner par groupes de trois ou quatre et disposent de deux heures. Afin d'assurer la dévolution du problème aux étudiants, l'assistant ne prend pas la responsabilité de la résolution du problème mais laisse les groupes d'étudiants explorer les pistes qui leur paraissent pertinentes et ce, pendant la première partie de la séance soit environ une demi-heure. Ensuite, dans un deuxième temps, l'assistant se permettra de fournir une aide technique ou des questions de relance aux étudiants qui seraient bloqués au démarrage, du type : "Quels peuvent-être les différents contenus ?", "Avec tel contenu, si on réalisait 10 tirages à quels résultats pourrait-on s'attendre ?", "Et avec 25 tirages ?". Des questions de

11. Cela peut être une boite contenant cinq cubes bleus ou rouges, un sac contenant des boules ou une bouteille et des billes, peu importe.

relance, de réflexion peuvent également être utilisées en cours de séance et seront fonction du type de raisonnement suivi par les groupes d'étudiants.

Si la situation initiale a les mêmes caractéristiques que celle proposée par Brousseau aux élèves de primaires, le format a changé puisqu'il s'agit ici d'une seule séance de 2h (au lieu d'une trentaine de séances de 10 minutes) dans laquelle les étudiants progressent en petits groupes (plutôt qu'une seul groupe classe) et seuls pendant une demi-heure, plutôt que suivis de près par l'enseignant qui, certes ne prend pas la responsabilité de la résolution du problème mais interagit avec les élèves et les aide à mettre à l'épreuve leurs conceptions. L'accompagnement diffère donc d'une part, à cause de contraintes techniques (le format) et d'autre part parce que nous voulions éviter que la séance soit un dialogue entre un seul étudiant et l'assistant, les autres étudiants restant passifs, en attente de la résolution du problème. En cela, travailler par petits groupes permet à un plus grand nombre d'étudiants de mettre leurs conceptions à l'épreuve et l'interaction entre les étudiants permet de devoir expliciter les raisonnements parfois implicites. Par ailleurs, la retenue de l'assistant à intervenir durant la première demi-heure a pour but de véritablement forcer les étudiants à se saisir du problème. Ceux-ci ne savent d'ailleurs pas que l'assistant les aidera plus loin dans la séance.

En termes de contrat didactique, l'idée est de partir sur le même schéma que celui utilisé lors de la première séance de travaux pratiques. Les étudiants avancent sur un problème général à leur rythme et par petits groupes, l'assistant passant entre les groupes pour suivre ou accompagner les étudiants dans leur raisonnement. Cette séance, comme les précédentes d'ailleurs, ne donne pas lieu à une évaluation : il n'y a ni test d'entrée, ni test de sortie, ni questions issues directement des travaux pratiques à l'examen. Les séances de travaux pratiques, par ailleurs non obligatoires, ont pour objectif d'aider les étudiants à comprendre la matière. Les étudiants peuvent donc se tromper durant cette séance puisqu'il n'y a aucune pénalité en cas de mauvaise réponse ou s'ils ne parvenaient pas à résoudre le problème au bout de la séance.

Raisonnements possibles

Le problème soumis aux étudiants est véritablement un problème ouvert : il n'y a pas une seule manière de l'aborder, ni une seule solution possible. Nous avons identifié quinze chemins possibles dont dix semblent accessibles aux étudiants. Nous proposons d'examiner en quoi ces approches diffèrent et, en particulier, de les caractériser à l'aide de trois critères : le modèle théorique sous-jacent, l'outil utilisé pour construire les distributions théoriques attendues et le critère d'arrêt utilisé pour déterminer la taille d'échantillon suffisante (voir figure 3.7). Nous avons également distingué les chemins que pourraient suivre des étudiants de ceux qui seraient suivis par des experts pour résoudre ce problème.

Du point de vue du **modèle théorique** sous-jacent, le modèle binomial est, d'un point

de vue théorique, parfaitement adapté à la situation puisqu'elle met en œuvre une répétition de n expériences indépendantes et dont la probabilité de chaque expérience individuelle est égale à π. Le fait que les étudiants aient manipulé ce modèle lors de la précédente séance de travaux pratiques et les difficultés potentielles liées à l'utilisation de distributions théoriques continues nous amènent à croire que les étudiants n'iront pas d'eux-mêmes vers d'autres modèles théoriques.

En ce qui concerne les **outils** qui seront manipulés, plusieurs possibilités s'offrent aux étudiants. Ils peuvent partir de tirages réalisés sur un modèle physique (une bouteille-modèle), ou bien réaliser des simulations dans un logiciel de type tableur, voire commencer avec un modèle physique et puis se diriger ensuite vers des simulations informatiques lorsque le besoin de grandes séries se fait sentir.

Notons que nous nous attendons à ce que la construction de simulations sur un tableur ne soit pas aisée pour ces étudiants. Or, nous souhaitons qu'ils puissent explorer cette voie sans être bloqués par le caractère technique de l'informatique. Nous avons donc pris le parti d'aider les groupes d'étudiants qui envisageraient cette voie à réaliser un tableau permettant de faire un grand nombre de tirages. Les assistants ont donc reçu comme consigne d'accompagner les étudiants dans la construction de cet outil si d'aventure ceux-ci s'engageaient sur cette voie.

A côté de ces deux outils qui ne nécessitent pas, en réalité, de connaissances formelles de la distribution binomiale, deux autres outils permettent également d'accéder aux distributions théoriques attendues : les tables statistiques et les formules de la fonction de probabilité binomiale. Ces deux outils ont été présentés et manipulés lors de la dernière séance de travaux pratiques donc il nous semble raisonnable de penser que les étudiants seront en mesure de les réutiliser dans ce contexte-ci. Cependant, ces outils ne sont pas des points de passage obligatoires puisque, nous l'avons vu, il est possible de démarrer par des tirages et des simulations de tirages.

Ces différents outils permettent globalement de construire la distribution des résultats attendus sous chacune des hypothèses. Ensuite, à partir de ces distributions théoriques, différents raisonnements sont possibles et **différents critères** peuvent être utilisés afin de déterminer le nombre de tirages nécessaires pour "connaître le contenu de la boîte".

La première démarche (voir figure 3.7) consisterait à associer à chaque résultat possible (donc l'ensemble des valeurs comprises entre 0 et 100 %) la conclusion que l'on tirerait. Par exemple, "si j'observe 0 % de boules blanches, je considère que la composition la plus vraisemblable est 5 boules noires" (proportion de blanches : $\pi = 0$). Entre 10 et 30 % de boules blanches, on pourra conclure que l'hypothèse la plus vraisemblable est $\pi = 0,2$; entre 30 et 50 %, on conclura que c'est $\pi = 0,4$ qui est la plus vraisemblable, etc. Ces intervalles sont assez intuitifs dans la mesure où ils encadrent simplement les probabilités attendues dans chacune des hypothèses. Par contre, ils ne sont pas optimaux car il faut prendre en compte le fait qu'un résultat ne peut

TABLE 3.8 – **Exemple de définition d'intervalles de décision à partir des résultats possibles d'un tirage**.

Plage de résultats	**Hypothèse considérée correcte**
$[0\ \% - 0\ \%]$	$\pi = 0$
$]0\ \% - 30\ \%[$	$\pi = 0,2$
$[30\ \% - 50\ \%[$	$\pi = 0,4$
$[50\ \% - 70\ \%[$	$\pi = 0,6$
$[70\ \% - 100\ \%[$	$\pi = 0,8$
$[100\ \% - 100\ \%]$	$\pi = 1$

entrer que dans un seul intervalle si on veut être en mesure de choisir une hypothèse à partir du résultat. Des intervalles du type : [10 %-29 %],[30 %-49 %],[50 %-69 %] et [70 %-89 %] seraient plus adaptés. De plus, ils laissent deux plages de résultat non couvertes par une décision : entre]0 %-10 %[et entre]90 %-100 %[. On pourrait donc éviter ces deux problèmes en utilisant les intervalles présentés dans le tableau 3.8.

A partir de tels intervalles, il est possible de réaliser des tirages sous différentes hypothèses et de voir, en fonction du nombre de tirages, comment évolue la proportion de conclusions correctes ou incorrectes (risque d'erreur). Les intervalles sont donc fixes, ils ne varient pas en fonction de la taille d'échantillon ; par contre, le risque d'erreur, lui, est variable en fonction de N. La démarche consiste donc à analyser la relation entre N et le risque d'erreur et à choisir un N qui donne un risque d'erreur acceptable.

La deuxième démarche consisterait à construire des intervalles en excluant les valeurs les moins vraisemblables sous chacune des hypothèses et puis à chercher la taille d'échantillon suffisante pour rendre ces intervalles disjoints (voir figure 3.7). En effet, on pourrait se demander : "Et si la composition de la bouteille était $\pi = 0,2$, quels résultats de tirage seraient possibles ? Si on prend n=10, par exemple, on constate qu'il est théoriquement possible d'observer n'importe lequel des résultats, on pourrait très bien observer 0 % de boules blanches comme 100 % : toutes ces valeurs sont *possibles*. Certaines sont, cependant, moins *probables* que d'autres. Observer 10 boules blanches d'affilée lorsqu'en réalité la composition de la bouteille est de $\pi = 0,2$ est possible mais hautement improbable. On peut donc décider de considérer que ce résultat (proportion observée de 100 %) est exclu sous $\pi = 0,2$. De manière plus générale, on peut décider que, pour chaque hypothèse, on exclut les 5 % des résultats les moins probables.

A côté de ces modèles, outils et démarches que les étudiants pourraient mettre en œuvre, on peut énoncer les modèles, outils et les démarches des experts.

Du point de vue des modèles théoriques, les experts pourraient utiliser la distribution bi-

nomiale ou bien être amenés à utiliser l'approximation normale de la distribution binomiale afin de réaliser ce calcul de taille d'échantillon. Bien qu'il s'agisse d'une approximation, elle est souvent réalisée en pratique et son impact devient négligeable lorsque la taille d'échantillon augmente. Ainsi, à partir d'un échantillon de 100 sujets, utiliser l'approximation normale au lieu des formules du modèle binomial ne change plus grand chose.

Les outils des experts seraient alors des simulations (réalisées avec un tableur ou avec un logiciel de statistique), l'utilisation directe de formules permettant de calculer un intervalle de confiance pour une proportion ou d'estimer une taille d'échantillon ou l'utilisation indirecte de ces mêmes formules via un logiciel de statistique.

A partir de simulations, en fixant une hypothèse comme hypothèse nulle et les autres comme hypothèses alternatives, il est possible de construire des intervalles de décision autour de l'hypothèse nulle et d'évaluer les qualités de ces intervalles à partir de différentes tailles d'échantillon en calculant la puissance du test pour différentes hypothèses. Cette fois-ci, le risque d'erreur associé à l'hypothèse nulle est fixe, il s'agit d'α fixé, par exemple, à 5 %, tandis que les risques d'erreurs liés aux autres hypothèses (β) varient en fonction de N. N est donc choisi de manière à garantir une certaine puissance pour discriminer deux hypothèses.

Un expert pourrait également partir directement d'une formule (ou d'un logiciel qui utilise une formule) pour estimer la taille de l'échantillon à partir des risques d'erreurs α et β associés aux hypothèses nulle et alternative, respectivement et à partir des hypothèses à départager, $\pi = 0,4$ et $\pi = 0,2$ par exemple.

Exemples de résolution

Les trois critères que nous venons de développer nous fournissent une grille de lecture pour différencier les approches du problème posé aux étudiants. Cette grille de lecture nous permet de mieux visualiser les différences entre le raisonnement attendu chez ces étudiants et le raisonnement suivi par les élèves de CM2 lors de l'expérience décrite par Brousseau (1974). Dans ce qui suit, nous allons montrer deux exemples de résolution du problème. Nous verrons d'abord la voie empruntée par les élèves dans l'expérience décrite par Brousseau puis un exemple de solution qui serait mise en œuvre par un statisticien expert. Cela nous amènera à voir en quoi ces démarches diffèrent du comportement attendu chez les étudiants.

Dans l'expérience de Brousseau, les élèves ont utilisé des tirages à partir d'un modèle physique (une bouteille dont on connait le contenu avec certitude, la bouteille Z), puis ont eu la possibilité de réaliser des simulations de tirages grâce à un logiciel informatique leur fournissant de grandes séries de résultats dans des conditions où la probabilité de la machine de hasard était fixée par eux.

Ensuite, à partir d'un jeu de devinettes, ils ont construit un ensemble de règles permettant de faire un pari sur la composition d'une bouteille à partir des résultats. Ces règles utilisaient des intervalles de décision. L'idée était que les tirages donnant une proportion observée entre 10 et 30 % conduisaient automatiquement à considérer $\pi = 0,2$ comme l'hypothèse correcte. Cet ensemble d'intervalles de décision a ensuite été affiné et éprouvé avec des séries de tirages (donnés par simulation) de différentes longueurs.

Les enfants ont donc pu constater qu'avec des séries de 15 tirages, on se trompe souvent (environ une fois sur deux) en utilisant ces intervalles tandis qu'avec 100 tirages, on se trompe beaucoup moins (environ une fois sur vingt). De là, ils pourraient donc conclure qu'une série de 100 tirages permettrait de connaitre le contenu de la bouteille avec suffisamment de certitude, avec un risque d'erreur tolérable.

Sur la figure 3.7, le chemin pris par ces élèves est dessiné en trait discontinu.

De leur côté, confrontés à ce même genre de problème, **les biostatisticiens** se baseraient certainement sur des outils et des démarches différentes. En effet, d'autres approches, plus directes, existent et sont utilisées pour répondre à ce genre de questions dans le domaine de la recherche.

Une approche consisterait à utiliser une formule permettant d'estimer une taille d'échantillon ou bien de passer par un logiciel ou un calculateur en ligne qui se base sur cette formule. Ce genre de formule nécessite de préciser plusieurs éléments :

1. Le nombre d'échantillons en jeu (classiquement un échantillon seul ou deux échantillons indépendants, deux groupes à comparer), ici il s'agit d'un test d'hypothèses impliquant un seul échantillon ;
2. Les deux hypothèses concurrentes, ici on pourrait prendre, par exemple, $\pi_0 = 0,4$ et $\pi_1 = 0,6$ [12], en fixant arbitrairement que H_1 correspond à $\pi_0 = 0,4$ et H_2 correspond à $\pi_1 = 0,6$ [13] ;
3. Le risque d'erreur α, placé ici à 5 % ;
4. Le risque d'erreur β, placé ici également à 5 %.

En prenant une formule basée sur l'approximation normale, la taille d'échantillon (n) serait obtenue de la manière suivante :

12. Prendre comme base les deux hypothèses les plus proches de la probabilité de 50 % permet de se placer dans le cas le plus défavorable, c'est-à-dire ici le plus variable, qui engendrera la taille d'échantillon la plus grande. A partir de cette valeur, on peut dire que si la taille d'échantillon est suffisante pour départager $\pi = 0,4$ et $\pi = 0,6$ entre elles, alors elle est aussi suffisante pour toutes les autres combinaisons d'hypothèses.

13. Nous aurions pu inverser H_1 et H_2 ; dans ce cas-ci, cela ne modifie pas la taille d'échantillon à laquelle nous serions arrivés.

$$n = \pi_0 * (1 - \pi_0) * (\frac{Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}}{\pi_0 - \pi_1})^2$$

$$n = 0,4 * (1 - 0,4) * (\frac{1,96 + 1,65}{0,4 - 0,6})^2$$

$$n = 78$$

Ou bien, en utilisant un logiciel passant par le modèle binomial, on aurait pour ces mêmes paramètres :

$$n = 84$$

(voir détails figure 3.8).

Comportements attendus chez les étudiants

Maintenant que nous avons décrit les raisonnements mis en œuvre par les élèves dans l'expérience de Brousseau (1974) ainsi que des exemples de raisonnements attendus chez les experts, nous pouvons décrire un exemple de ce que pourrait être la résolution du problème par les étudiants. Ce raisonnement attendu se base sur plusieurs hypothèses :

1. Face à une situation conçue pour faire émerger un raisonnement proche du test d'hypothèses, les étudiants seront en mesure de développer des connaissances qui s'approchent du savoir visé d'une manière similaire à celle développée par des élèves de CM2 dans l'expérience de Brousseau ;
2. Les étudiants sont capables de manipuler la distribution binomiale, c'est-à-dire qu'ils savent, au moins par groupe, reconstruire une distribution théorique binomiale. En effet, ce type de manipulations a été introduit et exercé lors de la précédente séance de travaux pratiques ;
3. Les éléments de contrat didactique concernant la façon de travailler, dans laquelle de petits groupes d'étudiants avancent à leur rythme sur un problème général, ne poseront pas de problème car ils auront été déjà mis en place lors de la première séance de travaux pratiques.

Un premier raisonnement consisterait à utiliser le modèle binomial ; par exemple, en suivant les étapes suivantes :

1. Connaitre le contenu = choisir entre six possibilités, à partir du résultat d'un tirage ;

2. Quels résultats pourrait-on observer avec $n = 10$? Même quand $\pi = 0,2$, on pourrait observer, par exemple, 4 boules noires (soit une proportion observée de 40 %) ou plus dans plus de 12 % des cas [14]. Avec $n = 10$, le choix que l'on pourrait poser à l'issue d'un tirage serait bien aléatoire (voir figure 3.9, haut) ;
3. Quels résultats pourrait-on observer avec $n = 100$? Cette fois, les résultats possibles dans les différents cas de figure se distinguent plus franchement (voir figure 3.9, milieu) ;
4. Pour faire un choix entre les hypothèses, il faudrait définir des zones de résultats qui nous amènent à choisir $\pi = 0,2$, $\pi = 0,4$, *etc.* On peut, par exemple, poser que, entre des proportions observées de 10 % et 30 %, on conclura que $\pi = 0,2$, entre 30 % et 50 %, on conclura que $\pi = 0,4$ et ainsi de suite.
5. Si on considère que $\pi = 0,2$ quand on observe entre 10 et 29 boules noires sur 100 tirages, alors la probabilité de se retrouver hors de ces limites quand $\pi = 0,2$ est de 1 %. Quand $\pi = 0,4$, alors la probabilité d'observer un résultat hors des bornes 30-49 est de 5 % (voir figure 3.9, bas),
6. On voit qu'avec $n = 100$, la probabilité de tirer la mauvaise conclusion est acceptable puisqu'elle est de l'ordre de 1 à 5 %, selon les cas de figure [15].
7. Si on prenait une taille d'échantillon plus importante, cette probabilité diminuerait encore.

Une deuxième manière de démarrer serait de réaliser des tirages. Par exemple, en suivant un raisonnement du type :

1. Réalisation de 10 tirages -> observation d'une certaine proportion de boules noires ;
2. En répétant plusieurs fois 10 tirages, on observe que les proportions varient fortement d'une fois à l'autre ;
3. Réalisation de séries de 50 tirages, les proportions sont moins variables ;
4. Il faudrait réaliser plus de tirages, et voir ce qui se passerait dans les différents cas de figure (différentes compositions possibles) ;
5. Construction de simulations sur un tableur pour décrire les variations possibles des résultats avec différentes tailles d'échantillon ;
6. A partir de là, on peut retomber sur le point 4. du raisonnement précédent.

Ces deux approches peuvent être en principe, du moins c'est l'hypothèse que nous formulons, suivies par des étudiants qui ignorent tout, au départ, du test d'hypothèses et des notions d'inférence statistique.

14. Les étudiants peuvent trouver les valeurs dans les tables statistiques, en appliquant la formule binomiale ou via un logiciel de type tableur.

15. Et de 0 % pour les hypothèses selon lesquelles toutes les boules sont du même type.

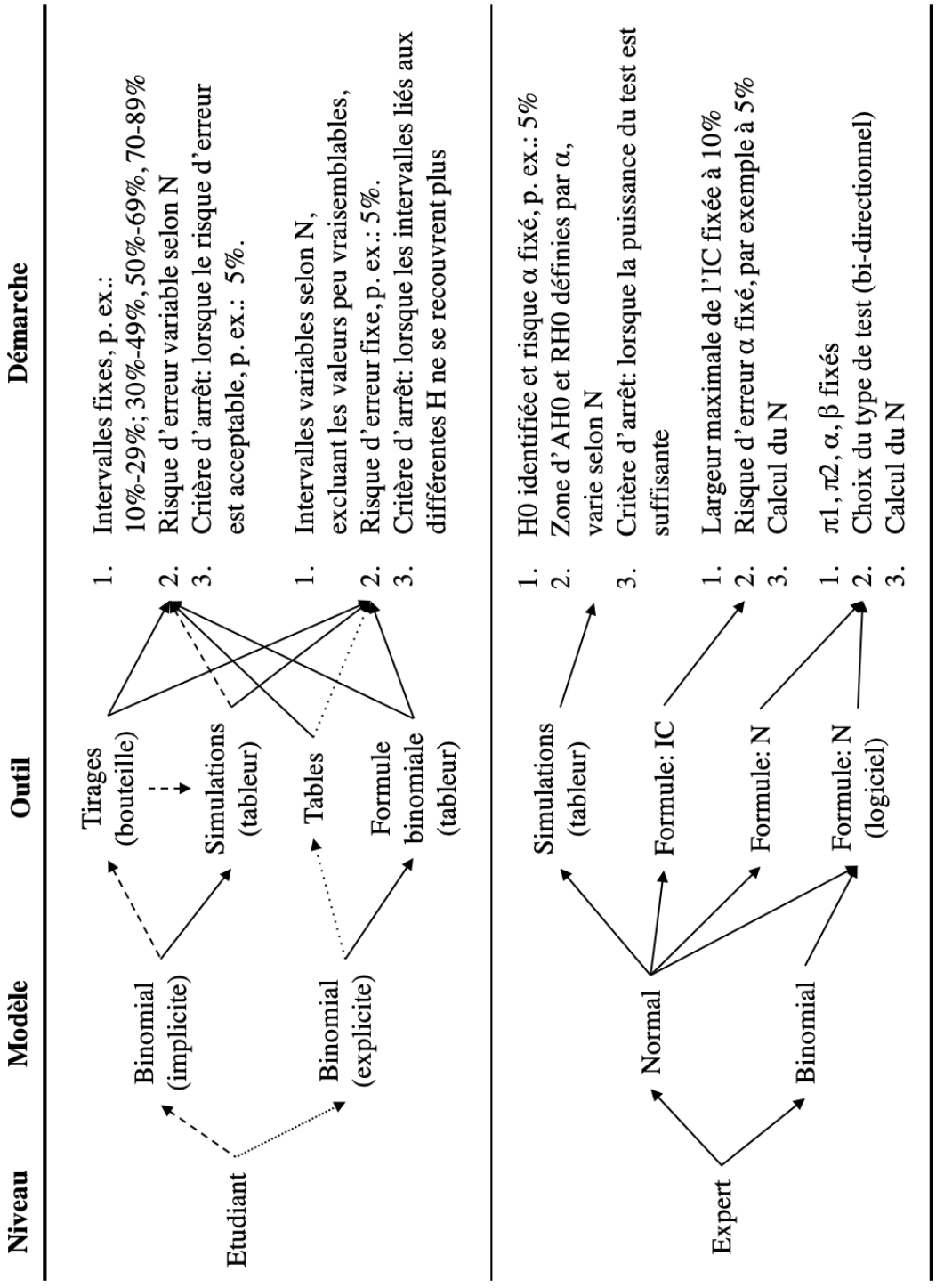


FIGURE 3.7 – **Caractéristiques des raisonnements possibles face à la situation fondamentale du test d'hypothèse**.

```
Exact - Proportion: Difference from constant (binomial test, one sample case)

Options:    α balancing: α/2 on each side

Analysis:   A priori: Compute required sample size
Input:      Tail(s)                    =   Two
            Effect size g              =   0,2
            α err prob                 =   0,05
            Power (1-β err prob)       =   0,95
            Constant proportion        =   0,4
Output:     Lower critical N           =   24,0000000
            Upper critical N           =   43,0000000
            Total sample size          =   84
            Actual power               =   0,9597606
            Actual α                   =   0,0444095
```

FIGURE 3.8 – **Capture d'écran du résultat d'une estimation de taille d'échantillon avec le logiciel G*Power.**

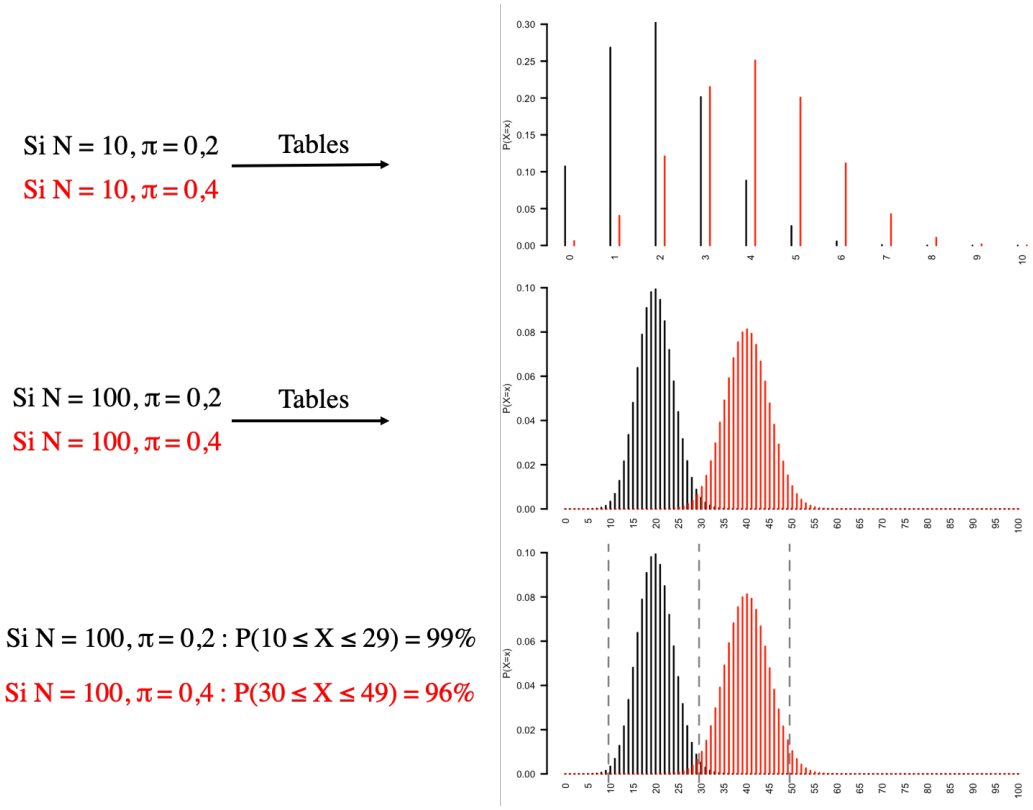


FIGURE 3.9 – **Exemple de résolution à partir du modèle binomial.**

3.3.4 Séance 4 : Institutionnalisation des connaissances

> "*Faire raconter aux enfants les évènements passés, et, sinon les propositions des uns ou des autres, du moins les résolutions retenues par l'ensemble, est une activité d'institutionnalisation indispensable aux élèves comme au maître. Non seulement elle permet de maintenir l'homogénéité (les élèves qui participent peu aux débats sont invités à les raconter) et de raviver l'intérêt des élèves (leurs remarques existent pour les autres), mais aussi elle favorise l'apprentissage. Elle demande un effort de formulation réflexive. Elle réorganise le passé et la mémoire de chaque élève (ce qu'il convient de retenir permet d'oublier en partie le reste). Elle permet aux élèves de décanter leurs questions et de faire de nouvelles propositions et au professeur de continuer son travail*" [Brousseau, 2005].

Durant la troisième séance, les étudiants avancent par groupes de trois ou quatre. Ils développent leur raisonnement et peuvent poser des choix différents, avoir des approches différentes d'un groupe à l'autre ou des justifications différentes des choix qu'ils prennent. Vers la fin de la troisième séance, ils reçoivent la consigne de se préparer à communiquer leur démarche aux autres groupes lors de la séance suivante. Ils devront résumer et expliquer leur démarche.

La quatrième séance sera donc une séance d'institutionnalisation et celle-ci est prévue pour être divisée en trois phases.

1. **Présentation des démarches suivies par les étudiants**. Chaque groupe d'étudiants présente et justifie sa démarche, avec ses mots.

   Les étudiants sont donc avertis lors de la troisième séance qu'ils auront à présenter leur réflexion au cours de cette séance-ci. Cela a pour but de les pousser à s'approprier le raisonnement, à le critiquer pour être capables de l'expliquer ;

2. **Reformulation des notions manipulées en termes de savoir savant**.

   L'assistant part d'une démarche présentée par un groupe d'étudiants pour décrire un raisonnement attendu en faisant référence aux notions de statistique visées : *test d'hypothèses, zone d'acceptation, risque d'erreur, hypothèses, distribution attendue sous une hypothèse, tirage, échantillon* ;

3. **Généralisation de la démarche**.

   Le raisonnement du test d'hypothèses est appliqué sur des exemples plus classiques, correspondant mieux à ce qui sera vu dans le cours théorique (et évalué).

3.4 Expérimentation

Dans cette section, nous allons décrire la mise au point du dispositif expérimental et les conditions de mise en œuvre de celui-ci.

Nous décrirons brièvement les observations réalisées lors des deux premières séances avant d'analyser, en deux temps, le déroulement de la troisième séance. Dans un premier temps, nous donnerons une description chronologique du déroulement des différentes séances. Pour chacun des quatre groupes d'étudiants, nous détaillerons ce qui nous apparait comme étant les principales étapes de leur raisonnement. Dans un second temps, nous compléterons ce tableau en présentant une analyse quantitative des concepts mobilisés par les étudiants durant leurs réflexions.

3.4.1 Mise au point

Le dispositif expérimental a d'abord été testé sur la cohorte d'étudiants en bachelier en sciences biomédicales et pharmacie. Le cours d'introduction aux notions de statistique médicale y est donné en 12 h et est placé dans les 60 crédits qui correspondent à la deuxième année d'études universitaires. Cette mise au point a eu lieu entre septembre et décembre 2018.

Cette cohorte d'étudiants est composée d'environ 200 étudiants répartis en 11 classes de travaux pratiques. La présence aux séances de travaux pratiques n'étant pas obligatoire pour ce cours, le nombre d'étudiants par groupe était d'une vingtaine au maximum.

Les travaux pratiques ont été donnés par plusieurs assistants différents dont l'un d'eux était l'investigateur de l'étude (B.B.). Les réunions de mise au point qui ont eu lieu entre les assistants ont permis de relever plusieurs points concernant le dispositif expérimental :

1. La première séance, consacrée à l'analyse descriptive, semble bien fonctionner dans le sens où les étudiants se saisissent du problème (comparer l'efficacité de trois régimes), investissent leurs conceptions, leurs manières de raisonner (utilisation de la moyenne, du minimum et du maximum) et en comprennent les limites ;
2. La deuxième séance est également assez satisfaisante pour l'équipe enseignante. Les étudiants avancent chacun à leur rythme mais semblent arriver, dans l'ensemble, à manipuler la distribution binomiale à travers les tables et les formules et à différencier un modèle et des observations expérimentales ;
3. La troisième séance est plus fastidieuse. Les étudiants ne semblent pas toujours très intéressés par le problème à résoudre et se désinvestissent rapidement du problème qui leur est posé. De leur côté, les assistants manquent de ressources pour relancer les groupes d'étudiants ayant du mal à entamer une réflexion sur le sujet. Au bout d'un moment,

les assistants doivent alimenter le raisonnement des étudiants et les mettre sur la voie ce qui, *de facto*, empêche la dévolution du problème aux étudiants ;

4. La quatrième séance se déroule généralement comme prévu, l'assistant faisant le lien entre les connaissances qui ont commencé à émerger lors de la troisième séance et les notions liées au test d'hypothèses.

Au vu de ces premières observations, il nous a semblé nécessaire d'opérer deux ajustements au dispositif expérimental.

D'une part nous voulions jouer sur la motivation des étudiants à résoudre le problème. A ce titre, il nous a semblé qu'intégrer la situation fondamentale au sein d'un contexte médical permettrait d'attiser leur curiosité et d'augmenter leur volition. Ils auraient ainsi plus l'impression de résoudre un problème qui les concerne et resteraient investis dans la tâche peut-être plus longtemps.

A ce titre, nous avons adapté les consignes de la manière suivante (voir figures 3.10, 3.11 et 3.12).

Par ailleurs, pour permettre aux assistants de mieux relancer les étudiants, nous avons travaillé ensemble différentes manières de résoudre le problème, explorant ainsi les différentes voies qui sont représentées sur la figure 3.7.

Surveillance de l'épidémie de grippe

Contexte

La grippe et le virus influenza.

"*Les virus influenza sont des virus à ARN, subdivisés en trois types différents (influenza A, influenza B et influenza C). Ce sont principalement les types A et B qui ont une importance clinique chez l'homme. (...) A l'échelle mondiale, le virus influenza est une cause importante de morbidité et mortalité. Globalement, en moyenne, 5 à 10 % des adultes et 20 à 30 % des enfants sont infectés par l'influenza chaque année, avec 3-5 millions de cas de maladie sévère et 1 million de décès par an* (WHO, 2012, Fischer et al., 2014). (...)

Chaque année, on observe une épidémie, habituellement durant l'hiver. Depuis 1918, quatre fois un nouveau virus influenza a fait son apparition et a provoqué une pandémie. La sévérité des épidémies et des pandémies dépend d'une part, des caractéristiques des virus circulant, et d'autre part, de la vulnérabilité de la population touchée. (...) (Sciensano, s.d.).

Composition du vaccin.

"*La composition des vaccins est revue chaque année sur la base des recommandations de l'Organisation Mondiale de la Santé (OMS). Pour la saison grippale 2018-2019, les vaccins quadrivalents contiennent les souches virales choisies en février 2018 sur la base des souches qui ont dominé lors de la saison précédente." (Sciensano, s.d.).*

Cette année, les vaccins ciblent les souches A H1N1, A H3N2, B/Victoria, B/Yamagata (Sciensano, s.d.). L'efficacité de ce type de vaccin envers les souches A H1N1, B/Victoria et B/Yamagata semble modérée, tandis qu'elle est faible ou nulle envers la souche H3N2 (Darvishian et al., 2017).

FIGURE 3.10 – **Consignes adaptées pour la troisième séance de travaux pratiques** (page 1/3)

.

Surveillance de l'épidémie

Au vu de l'impact de la grippe sur la population, un système de surveillance de l'épidémie et des souches circulante a été mis en place. Il est composé de médecins généralistes " vigies " qui envoient des frottis nasopharyngés de patients souffrant d'états grippaux à un laboratoire qui en détermine le sous-type viral.

Le suivi des sous-types circulants permet de :

- *prédire l'efficacité vaccinale* ;
- *mesurer la sensibilité des souches circulantes aux antiviraux* ;
- *de composer les vaccins pour la saison prochaine* ;
- *de donner l'alerte en cas de pandémie.*

Références :

*D*arvishian M, Dijkstra F, van Doorn E, Bijlsma MJ, Donker GA, de Lange MMA, et al. (2017) Influenza Vaccine Effectiveness in the Netherlands from 2003/2004 through 2013/2014 : The Importance of Circulating Influenza Virus Types and Subtypes. PLoS ONE 12(1) : e0169528. doi :10.1371/journal.pone.0169528

Fischer WA 2nd, Gong M, Bhagwanjee S, Sevransky J. Global burden of influenza as a cause of cardiopulmonary morbidity and mortality. Glob Heart. 2014 Sep ;9(3) :325-36. doi : 10.1016/j.gheart.2014.08.004. Epub 2014 Oct 31.

Sciensano (s.d.) – Epidémiologie des maladies infectieuses – Influenza - https ://epidemio.wiv-isp.be/ID/diseases/Pages/Influenza.aspx, consulté le 14/2/2019.

World Health Organisation (WHO). Vaccines against influenza WHO position paper - November 2012. Wkly Epidemiol Rec. 2012 Nov 23 ; 87(47) : 461-76.

FIGURE 3.11 – **Consignes adaptées pour la troisième séance de travaux pratiques** (page 2/3)

.

Mise en situation

*V*ous faites partie du centre chargé de surveiller les sous-types circulant dans la population belge durant l'épidémie de la saison 2018-2019. En particulier, vous devez déterminer quelle est, cette année, la proportion de sous-types A H3N2 (contre lequel les vaccins ont eu très peu d'efficacité jusqu'à présent).

[Tirage aléatoire de 5 sous-types parmi un ensemble contenant des sous-types A H3N2 et des autres sous-types].

La question à laquelle vous devez répondre est la suivante :

Combien de prélèvements faudrait-t-il analyser pour connaître la proportion de virus A H3N2 durant l'épidémie de grippe de cette année ? 10, 20, 50, 100 ou 200 ?

Consignes

*P*our répondre à cette question, vous travaillerez par petits groupes, en autonomie. (Vous pourrez toutefois faire appel à un consultant externe pour vous débloquer en cas de blocage ;-)

Une fois que vous aurez déterminé le nombre de prélèvements à réaliser, expliquez votre démarche (dans un langage simple) afin de justifier le nombre de prélèvements que vous demandez à votre hiérarchie.

Enfin, réalisez les tirages et décrivez les résultats obtenus et les limites de la méthode utilisée.

FIGURE 3.12 – **Consignes adaptées pour la troisième séance de travaux pratiques** (page 3/3). La phrase entre crochet signifie que la composition (à l'aveugle) du contenu de la boîte se fait en début de séance, par tirage aléatoire de cinq balles dans un ensemble comprenant des balles représentant des virus de sous-type H3N2 et des balles représentant des virus d'autre sous-types.

3.4.2 Mise en œuvre

Ce dispositif expérimental adapté (voir figures 3.10, 3.11 et 3.12) a donc été utilisé sur une deuxième cohorte d'étudiants bacheliers en médecine. Ces étudiants étaient pour la plupart dans leur première année d'études universitaires car le cours d'introduction aux notions de statistique médicale fait partie des 60 premiers crédits ECTS du bachelier. Cette cohorte comportait également environ 200 étudiants divisés en onze classes d'une vingtaine d'étudiants maximum. A nouveau, les séances de travaux pratiques n'étaient pas obligatoires et précédaient la matière vue en cours. L'expérience s'est déroulée entre les mois de février et juin 2019.

Cette fois, l'investigateur ne faisait pas partie de l'équipe d'assistants assurant les séances de travaux pratiques.

Les observations concernant le déroulement des séances de travaux pratiques sont issues d'entretiens avec les assistants, sauf pour la troisième séance, basée sur une situation fondamentale pour le test d'hypothèse, pour laquelle des enregistrements ont été réalisés.

Parmi les onze classes, quatre ont été sélectionnées pour un enregistrement lors de la troisième séance. Ainsi, au début de la séance, l'investigateur (présenté comme un chercheur, menant une thèse sur l'enseignement de l'inférence statistique), se présente et sélectionne aléatoirement un groupe d'étudiants dans la classe.

Ce groupe est amené dans un local à part afin que l'enregistrement (sonore uniquement) des raisonnements d'étudiants soit de qualité suffisante. Les étudiants étaient avertis que leurs discours seraient enregistrés et utilisé à des fins de recherche de façon anonyme. A aucun moment, l'investigateur n'a demandé ni cherché à connaitre l'identité des étudiants concernés.

Les étudiants enregistrés sont installés dans le local, disposent de la feuille des consignes, d'une boite contenant 5 boules, certaines étant marquées (représentant la souche virale B) d'autre n'étant pas marquées (représentant la souche virale A). Ils peuvent l'utiliser pour faire des tirages, avec remise, d'une boule à la fois avec remise. Ils n'ont pas le droit d'ouvrir la boite.

Au départ, l'investigateur leur donne la feuille de consigne ainsi que la boite et puis laisse les étudiants seuls pendant plusieurs minutes pour lire à leur aise les consignes et démarrer leur réflexion. Ensuite, régulièrement, l'assistant revient pour demander aux étudiants d'expliciter leur raisonnement, leur démarche. Son objectif est d'assurer la dévolution du problème aux étudiants, il s'abstient donc de favoriser une voie plutôt qu'une autre ou de valider/infirmer les raisonnements des étudiants. Cette attitude neutre sera maintenue le plus longtemps possible afin d'éviter que les étudiants ne se désinvestissent du problème.

3.4.3 Séances 1 et 2

Globalement le retour de l'équipe d'assistants concernant le déroulement des deux premières séances est plutôt positif.

Lors de la première séance, les étudiants se sont engagés dans la tâche qui leur était confiée, ont tenté de résoudre le problème avec leurs connaissances antérieures et puis, voyant que cela ne suffisait pas, se sont mis à utiliser des notions statistiques nouvelles. A mi-séance environ, l'assistant a fait le point sur les notions utilisées dans les différents groupes pour décrire les données à l'aide de statistiques en faisant le lien avec les notions théoriques du cours de statistique. De même, vers la fin de la séance, l'assistant a fait le point sur les représentations graphiques des distributions expérimentales. Dans certaines classes, cette mise au point a été l'occasion, sur base d'un exemple concret, de fournir des éléments de réponse à la question "peut-on faire dire ce que l'on veut à une analyse statistique ?".

La deuxième séance s'est également déroulée sans accrocs majeurs, les étudiants et les assistants ne semblant pas du tout perturbés par cette manière de travailler, somme toute assez classique, puisqu'il s'agissait de suivre une séquence didactique assez structurée. A certaines étapes, les assistants relatent des différences de rythme entre les étudiants mais, dans l'ensemble, les étudiants semblent arriver au bout de ce qui est demandé dans les temps impartis. Lors de cette séance-ci, l'assistant a également fait le point à différents moments. Au cours de cette séance, la notion de modèle théorique n'a pas semblé poser de difficulté majeures aux étudiants. Cette séance était également l'occasion de montrer, dans un contexte où le modèle est *a priori* très proche de la réalité (modèle binomial et lancers de pièces de monnaie), d'une part, que des écarts entre des observations et un modèle pouvait avoir lieu par hasard et, d'autre part, que ces écarts avaient tendance à se réduire lorsque le nombre de tirage augmentait. Ces observations n'ont pas semblé poser de difficulté aux étudiants. Par ailleurs, les étudiants se sont également montrés capables de manipuler la distribution binomiale à l'aide des formules et des tables une fois que cela leur avait été expliqué.

3.4.4 Séance 3 : description chronologique

Groupe 1

Dans les quatre prochaines sections nous allons tenter de décrire, pour chacun des quatre groupes enregistrés, le déroulement de la troisième séance de travaux pratiques. L'idée est de montrer d'une part, les étapes du raisonnement de chaque groupe d'étudiants et d'autre part, de fournir la matière qui servira à l'analyse *a posteriori*. Toutefois, vu la longueur des discussions, nous avons jugé opportun de ne retranscrire ci-dessous que les éléments qui nous semblaient les plus susceptibles d'être exploités dans la suite de l'analyse.

Dans ce qui suit, les neuf étudiants répartis en quatre groupe sont notés E1 à E9, l'investigateur (B.B.) est noté I. Le texte entre parenthèses décrit une action ou la réaction d'un intervenant à ce que dit l'émetteur du message. Les raisonnements des étudiants sont retranscrits sans modification (et donc sans correction). Les erreurs de français des étudiants et de l'investigateur n'ont pas été corrigées lors de la retranscription.

1. **Éléments de clarification**

   Dans un premier temps, les deux étudiantes commencent par essayer d'identifier les données à disposition. Puis elles reformulent la question. Elles poursuivent avec des questions de clarification et donnent ensuite une réponse intuitive à la question posée.

   E1 : "*Mais non, il n'y a pas de données*"

   E2 : "*Donc là on doit inventer des données ?*"

   E1 : "*Bin, les variables qu'on doit étudier c'est juste le nombre de prélèvements et le type de sous-unités alors.*"

   E1 : "*Donc, en gros, on cherche le nombre de tirages qu'on doit faire pour analyser directement les proportions*"

   E1 : "*On n'a pas du tout les (. . .) proportions entre les deux virus alors ?*"

   I : "*Non, parce que l'idée justement, c'est de pouvoir le trouver.*"

   E1 : "*Ah, ok.*"

   E1 : "*Les deux virus, on ne peut pas les avoir en même temps je suppose ?*"

   I : "*On va dire que non.*"

   E2 : "*10 ça paraît peu, non ?*"

2. **Approche par l'analyse combinatoire**

   Ensuite, rapprochant cette situation didactique d'un problème d'analyse combinatoire (abordé lors de la précédente séance de travaux pratiques ou dans l'enseignement secondaire), les étudiantes tentent, sans succès, de réaliser des calculs à partir des données qu'elles ont identifiées.

   E1 : "*On l'a fait la fois dernière avec le tirage de pièces.*"

   E2 : "*Je ne vois pas comment arriver à trouver ça*".

   E1 : "*J'ai vu une loi l'année dernière où on prenait des chiffres de tel à tel nombre et qu'on pouvait les prendre plusieurs fois ou indépendamment mais je ne me souviens plus de la formule.*"

   E1 : "*Combien de jeux différents avec 32 cartes, c'est un peu ça, c'est combien de séries on doit faire avec trois chiffres.*"

   E1 : "*Ca veut dire que si le médecin (. . .) prend 10 individus, il peut envisager 47 possibilités, donc . . .*"

   E2 : "*J'ai calculé pour 10 ça fait 192. Mais en faisant ça, je ne sais pas comment on va trouver*".

   E2 : "*Non, je ne sais pas comment faire.*"

   Au bout d'une trentaine de minutes (29 :15), l'investigateur intervient pour tenter de diriger les étudiantes vers une piste plus fructueuse. Il tente de reformuler la question. Il essaie ensuite de fermer la piste de l'analyse combinatoire.

I : "*Ici (...) il y a cinq boules, cela veut dire qu'il y a quoi comme possibilités ?*"

I : "*Si on prend 10 tirages, est-ce qu'on sait déterminer le contenu de la boite avec assez de certitude ou est-ce qu'il faut plutôt 20 tirages pour [le] déterminer ?*"

E2 : "*Moi 10, ça me paraît peu.*"

I : "*Comme vous vous en doutez, plus on va prendre de patients, plus ce sera précis, mais on ne peut pas se permettre de prendre 1000 patients si en fait avec 10 c'était assez (...) Donc on sait bien que plus c'est toujours mieux mais il faut savoir quand est-ce que (sic) c'est déjà suffisant.*"

E1 : "*Sur les 10 patients, tu as 47 possibilités mais sur l'échantillon que tu as pris tu n'as pas observé les 47*".

I : "*En fait le 47, ça ne représente pas exactement ce que vous pensez, je crois*".

E1 : "*Et si on tire 4 malades du premier et puis 6 malades du 2e (sous-type) ou qu'on tire 6 malades du deuxième et puis 4 malades du premier c'est la même chose ? C'est la même proportion finalement ?*"

E1 : "*Je me souviens m'être posée toutes ces questions-là l'année dernière, il faut que les formules reviennent*".

I : "*Essayez un peu de voir (...) quelles sont les combinaisons possibles.*"

I : "*Ok, je vois, tu as fait les combinaisons possibles de résultat de 10 tirages, mais quelles sont les combinaisons possibles de ce qu'il y a là-dedans, sachant que là-dedans il n'y a que cinq objets*".

I : "*Si je fais un seul tirage, est-ce que je peux dire, à bin voilà c'est une H1N1, du coup il n'y a que des H1N1 ?*"

E2 : "*Non c'est pas assez.*"

I : "*Et si j'en fais 5, et que j'ai à chaque fois (...) des H1N1, est ce que je peux dire qu'il n'y a que des H1N1 ?*"

E2 : "*Mais je ne sais pas comment faire pour trouver*"

3. **Sur la piste de la distribution binomiale**

44'30 : face au blocage, qui commence à devenir gênant, l'investigateur oriente les étudiantes plus directement vers l'identification des différentes hypothèses puis la construction de la distribution attendue.

I : "*Si on fait 10 tirages et que c'est ça la composition, je m'attendrais à quel genre de résultats ?. Quels seraient les résultats probables et quels seraient les résultats difficiles à obtenir ?*"

E1 : "*Ce n'est pas la même chose que de trouver le nombre de combinaisons qu'il faut faire pour avoir envisagé toutes les possibilités alors ?*"

I : "*Non, ici on dit : si c'était cette composition-là qui était la bonne, et si, ça fait déjà deux si, et si on prenait dix tirages.*"

E1 : "*Donc une combinaison, et la combinaison c'est les pourcentages ?*"

I : "*Oui, c'est 0% de H3N2 ou 20% de H3N2 ou 40% de H3N2, ou 60 ou 80 ou 100. Ici on sait que ce n'est pas autre chose que ces (six) valeurs-là parce que je n'ai mis que 5 balles.*"

E1 : "*Donc si on imagine qu'il y a 1/5e qui est H3N2 et 4/5e qui est H1N1 ? (oui) Et le deuxième si c'est le nombre de tirages ? (oui, si on en prend 10, par exemple).*"

E [16] (soupir)

I : "*Est-ce que vous ne voyez pas une manière de savoir quels sont les résultats, quelle est la*

16. E1 ou E2, impossible dans ce cas de les identifier sur base de l'enregistrement audio.

probabilité de chaque résultat si on avait X tirages, ou N tirages ?"

E1 : "*Bin, c'est des tirages où on remet à chaque fois tout dans la boite ?*

I : "*Oui.*"

E1 : "*Bin, on aura en grande majorité ceux-là et ça c'est difficile à obtenir.*"

I : "*Oui, ça c'est au niveau qualitatif (. . .) mais est-ce qu'on ne sait pas le mesurer ?*"

E2 : "*C'est un truc de probabilités mais je ne sais pas c'est lequel*".

Les étudiantes repartent vers les notions de probabilité et d'analyse combinatoire. L'investigateur fait maintenant un lien explicite au contenu de la séance précédente et aux outils à utiliser.

E1 : "*C'est deux évènements qui sont indépendants, puisqu'on ne peut pas observer les deux ensemble (H3N3 et H1N1 en même temps). Donc il n'y a pas d'intersection là. Mais c'est la probabilité d'obtenir... On doit tout faire (toutes les combinaisons possibles). 1/5e puis l'autre 4/5e, 4/5e, 4/5e, 4/5e.*"

I : "*Il y a peut-être un moyen de calculer les probabilités dans ce cadre-là. Vous avez pas vu, au TP précédent, quelque chose qui vous dit, tout seul, les probabilités de chaque résultat ? Non ?*"

E : "*Je ne me souviens plus*"

I : "*Vous aviez fait là, avec les microsatellites, vous avez lancé des pièces (oui, ça oui) et puis ça avait fait une distribution expérimentale. Vous aviez vu les résultats. Et puis vous aviez aussi une distribution théorique, que vous aviez faites avec un arbre de probabilité ? (oui) Et après, il y avait un exercice où on généralisait... Et donc il y a une méthode qui permet de connaitre les probabilités...*

E1 : "*Ah oui, c'est avec les tables, les deux tables (binomiale et Poisson). Oui, je vois, je ne sais plus laquelle des deux correspond à quoi. Il y en a une, la binomiale, c'est avec le nombre, il faut déterminer un échec et un . . . truc.*"

E2 : "*Bin là, en soi, c'est pas Poisson, parce que binomiale c'est échec et succès mais là c'est pas un échec, c'est pas un succès, non ?*"

E1 : "*Bin si, car tu peux considérer que, si tu tires quelqu'un qui a le deuxième virus... Parce que l'autre, dans l'autre cas, dans la loi de Poisson je pense que c'est quelque chose d'infini là*".

E2 : "*Donc on va devoir regarder par rapport aux tables et tout, alors ?*"

E1 : "*Bin la table, elle va nous donner la probabilité d'avoir autant, beaucoup de ça*"

E2 : "*Il nous a fallu du temps...*"

Comment utiliser les tables ? Une des deux étudiantes réexplique à l'autre ce que représentent les paramètres n et π de la distribution binomiale. Pendant plusieurs minutes, les étudiantes cherchent comment utiliser cette table (l'investigateur n'est pas dans la pièce).

E1 : "*donc... probabilité... binomiale de paramètre (pi) 0,2 et le tirage... 5 tirages ici ? On est toujours dans la boite ? On n'est pas encore dans les maladies.*"

E2 : "*Et du coup là quand on fera avec ça, à la place de mettre 5, on met genre 10, 20, 50,*

100, 200 ? (l'autre étudiante acquiesce) Ah, à l'aise, c'était long à trouver".

E1 : "*Et on devra trouver des proportions sur 10*"

E2 : "*Et tu dois faire avec 0.2, enfin 1/5, 2/5, 3/5, 4/5 et 5/5 ? Tu ne fais pas juste avec 1/5.* "

E2 : "*Maintenant il faut interpréter le chiffre*"

E2 : "*Et du coup, à chaque fois qu'on va regarder, genre là pour 10 tirages, à chaque fois (pour chaque pi) on va arriver à 1, la probabilité d'avoir ça c'est 100%* "

E1 : "*Oui, mais il faut faire les différences dans les colonnes, comme on a fait la semaine dernière (...) on doit faire cette colonne (ligne)-là moins la colonne d'en dessous.*"

Les étudiantes essaient ensuite de combiner les différentes informations pour en faire un calcul.

E1 : "*Donc on va avoir 0,9803-0,0219, ça veut dire qu'on a 5% (de chances) d'avoir (HN03 ? Rires hoo ça va, la chimie me hante). Donc on a été dans la table, on a été chercher l'intersection 5e ligne, puisque 5 tirages...* "

E2 : "*Et du coup, tu ferais un tableau Excell avec ça, genre parce qu'on doit faire avec 0,2, enfin 1/5, 2/5, 3/5 ? Ou on fait juste avec 1/5 ? Il faut faire avec tout !*"

E1 : "*Mais là on a déjà des pourcentages, attends, une chose à la fois*"

4. **Rappel concernant le fonctionnement de la table de la distribution binomiale**

59'02 Une deuxième fois, l'investigateur quitte son rôle d'observateur pour prendre un rôle d'enseignant afin de rappeler comment lire les tables statistiques. La raison de ce changement de rôle, qui a déjà commencé lorsque l'investigateur a mis les étudiantes sur la voie de la distribution binomiale, est double. D'une part, il s'agit de permettre aux étudiantes d'atteindre une résolution du problème dans le temps imparti et de ne pas les laisser durant 2h sur de mauvaises pistes. D'autre part, cela permet à l'investigateur de recueillir les réactions de ces étudiantes à la logique du test d'hypothèses.

I : "*En fait, la table il faut peut-être la lire un peu autrement. Ça, ça veut dire : si on fait 10 tirages et si la proportion de H3N2 est de autant, 1/5, alors avoir, sur les 10 patients, 0 qui est H3N2, ça se peut, il y a 10% de chances que cela arrive. En avoir 0 ou 1, ça se peut même un peu plus, il y a 37% de chances que cela arrive*".

5. **Les résultats rares sous une hypothèse la contredisent**

Ensuite, l'investigateur explique aux étudiantes que, sous une certaine hypothèse, les résultats rares remettent en question celle-ci, permettent d'une certaine manière d'exclure une hypothèse. Par exemple, observer 10 succès en 10 tirages permet d'exclure que la probabilité de succès est de 20 %.

Les échanges qui suivent indiquent que cette manière de procéder est loin d'être intuitive pour ces étudiantes qui en arrivent parfois à affirmer des choses en dépit du bon sens.

E1 : "*Cela veut dire que, plus on augmente le nombre de personnes, plus il y a des combinaisons qui sont rares*"

E2 : "*Du coup, le mieux c'est d'avoir moins de personnes*"

E1 : "*(pas forcément car) si tu as plus de personnes, tu peux éjecter plus facilement les pourcentages*".

6. **Si les résultats sont compatibles avec plusieurs hypothèses, il n'est pas possible de tirer une conclusion**

   Par la suite (01 :15 :51), l'investigateur amène l'idée d'intervalles qui se recouvrent.

I : "*Si tous les résultats qu'on pourrait observer nous permettent de dire clairement 'c'est plutôt (cette composition-ci) ou plutôt celle-là qui est bonne' alors on est bon, on a assez de tirages. Si, par contre, il y a des résultats qui sont (. . .) compatibles avec plusieurs hypothèses, alors on ne sait pas les discriminer.*"

E2 : "*Bin du coup, ça veut dire qu'on doit prendre moins de monde ?*"

E1 : "*Je suis perdue*"

7. **Résumé final**

   Après que l'investigateur a réexpliqué la logique, les étudiantes ont quelques minutes pour tenter de faire une synthèse du raisonnement qui a été suivi (et proposé par l'investigateur). Le vocabulaire choisi semble démontrer que l'étudiante applique une recette de cuisine sans véritablement en comprendre la logique.

E1 : "*Et bien on a choisi une. . . on a décidé d'imposer une probabilité, enfin une proportion de malades dans la population. Et puis on a regardé quelle était la chance entre guillemets d'avoir 0 malade, 1 malade, 2 malades pour pouvoir essayer d'estimer les probabilités qui seraient les plus acceptables. On a fait ça pour différentes tables, de n différents pour essayer de voir si, pour essayer d'éliminer déjà des probabilités : 1/5, 2/5, etc. Et puis on a regardé les fourchettes de nombre de gens qui sont malades dans la population qui étaient les plus probables, en fonction des probabilités qu'on trouvait. Avec ces fourchettes-là, on a essayé, entre les différentes colonnes, donc entre les différentes probabilités, de regarder, enfin d'essayer d'avoir deux zones, radicalement entre guillemets, différentes pour pouvoir, en fonction du résultat, déterminer dans quelle zone est-ce qu'on se situe. Et donc, en fonction, trouver la probabilité. Et ensuite, donc on a déjà trouvé que pour 50 personnes, il était déjà possible de séparer deux fréquences, enfin deux pourcentages entre guillemets, si on admettait un risque d'être en dehors de la zone de 10%. Et en fonction du risque que l'on admet, et bien il faudrait plus ou moins de tests, de gens à tester.*"

Groupe 2

1. **Réponse intuitive et décodage du contexte**

ANNEXE I : TABLES STATISTIQUES

Table des distributions Binomiales

P(X≤x)

X= nombre de succès

Π = probabilité de succès

N= nombre de réalisation de l'épreuve

n = 10

n= 10

		π 0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
x	**0**	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,9139	0,7361	0,5443	0,3758	0,2440	0,1493	0,0860	0,0464	0,0233	0,0107
	2	0,9885	0,9298	0,8202	0,6778	0,5256	0,3828	0,2616	0,1673	0,0996	0,0547
	3	0,9990	0,9872	0,9500	0,8791	0,7759	0,6496	0,5138	0,3823	0,2660	0,1719
	4	0,9999	0,9984	0,9901	0,9672	0,9219	0,8497	0,7515	0,6331	0,5044	0,3770
	5	1,0000	0,9999	0,9986	0,9936	0,9803	0,9527	0,9051	0,8338	0,7384	0,6230
	6	1,0000	1,0000	0,9999	0,9991	0,9965	0,9894	0,9740	0,9452	0,8980	0,8281
	7	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984	0,9952	0,9877	0,9726	0,9453
	8	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9983	0,9955	0,9893
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9990
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

FIGURE 3.13 – Exemple de la table statistique donnant la fonction de répartition pour la distribution binomiale

Dans un premier temps, les étudiants fournissent des réponses intuitives tout à fait pertinentes. Ensuite, ils parviennent assez bien à distinguer l'essentiel de l'accessoire dans le contexte et identifient, notamment, le rôle non essentiel du contexte médical dans la mise en situation. La formulation, dans leurs termes, de la question globale semble montrer qu'ils ont assez bien compris ce que l'on attendait d'eux.

E3 : "*Mais en vrai, moi, j'aurais d'office dit 200 car plus il y en a, plus c'est susceptible de t'aider non ?*"

E5 : "*Oui, mais ça ne doit pas être aussi simple*"

E5 : "*L'efficacité de ce type de vaccin envers les souches (. . .) semble modérée tandis qu'elle est faible ou nulle envers la souche H3N2.*"

E3 : "*Oui mais ça, ça a un rapport ?*"

E5 : "*Je n'en sais rien, pourquoi ils le mettent ?*"

E3 : "*Mais c'est pour te mettre en situation non ?*"

E5 : "*Oui*".

E3 : "*Donc, en gros, ils t'envoient des trucs de patients en état grippal, et puis après, toi, tu dois déterminer le sous-type et donc là avec combien de frottis tu penses pouvoir déterminer la proportion de H3N2 dans l'épidémie totale de grippe cette année, c'est ça ?*"

E3 : "*je ne sais pas, j'ai tellement lu dans le livre qu'il est écrit qu'il faut que l'échantillon soit grand qu'on va mettre 200 ?*"

E4 : "*Moi aussi, je suis d'accord pour faire*

plus d'échantillons mais je me pose la question, je cherche dans le texte s'il n'y a pas une contrainte en te disant, je ne sais pas, que ça coûte cher ou quoi que ce soit... (rires) Non, non, mais parfois tu peux avoir une petite contrainte quoi, que le nombre d'échantillons est limité ou que sais-je..."

Au bout de 10', l'investigateur confirme l'idée pressentie par les étudiants qu'un échantillon plus grand est effectivement mieux mais qu'il y a des contraintes qui poussent à en prendre le moins possible également. Il confirme également le rôle accessoire du contexte.

E3 : "*Mais est-ce que ça change vraiment quelque chose qu'on nous dise, enfin, ce qu'on nous dit par rapport à l'efficacité du vaccin ou pas ? Ou c'est du blabla pour la mise en situation ?*"

I : "*C'est plutôt de la mise en situation.*"

2. **Identification d'un problème dans le contexte**

Un étudiant utilise les éléments du contexte pour élaborer un raisonnement statistique complexe et pertinent.

E4 : "*Oui, mais on dit que l'efficacité du vaccin de cette année a bien réussi, par exemple, pour les gens qui ont eu la grippe type Victoria ou Yamagata par exemple (l'investigateur acquiesce). Et on dit aussi, dans le contexte, que ceux qui ont eu la grippe H1N1 et Victoria ou Yamagata semblent modérés, cela veut dire que ça a plus ou moins fonctionné le vaccin pour ces gens-là. Mais, par contre, on nous dit, c'est pour ça que ça m'a intéressé, on nous dit que tandis qu'elle est faible ou nulle envers les souches H3N2. Donc ça veut dire que, quelque part, la majorité des gens chez qui on va faire un prélèvement ont de fortes chances d'avoir cette souche-là.*"

I : "*Ha oui, je vois, dans ce sens là. Ça c'est si tout le monde est vacciné. Si les gens étaient tous vaccinés, effectivement, et que l'efficacité était de 100% pour les H1N1 alors tu as raison que, dans l'échantillon, on ne trouverait plus que des H3N2.*"

E4 : "*Je sais qu'on trouvera des H1N1 et des Victoria mais c'est pour justifier qu'on prenne un échantillon de 200 personnes, pour justifier qu'on n'est pas obligés de prendre 2000 personnes.*"

I : "*Et comment tu justifies 200 plutôt que 2000 ?*"

E4 : "*Et bien vu qu'on sait que la majorité des gens auront la souche H3N2 et pas les autres, vu que le vaccin a plus ou moins fonctionné*".

E3 : "*Oui mais ça, ça ne va pas te donner une proportion dans la population justement, ta proportion elle sera biaisée si tu fais ça. Parce que, justement, tu auras beaucoup d'H3N2 vu que le vaccin pour les autres (souches) a marché. Alors que justement ce que tu veux c'est avoir une proportion normale. Tu vois, si tu disais que 100% des gens étaient vaccinés, et qu'après tu cherchais la proportion d'H3N2, tu aurais d'office plus d'H3N2 vu que tout le monde a été vacciné et que le vaccin est efficace pour ceux qui ont H3N1. Si tu prends que des frottis de gens malades, vu que c'est le vaccin pour H3N2 qui ne marche pas bien, tu aurais d'office beaucoup d'H3N2 alors que là, nous ce qu'on veut c'est la proportion normale si personne n'était vacciné, tu vois ? Enfin les*

TABLE 3.9 – **Proportions infectées vs proportions symptomatiques à un temps t**. Dans cet exemple, on fait l'hypothèse que tous les prélèvements sont réalisés au même moment et que les individus peuvent se répartir en trois classes exclusives : non-infecté, infecté par la souche A ou infecté par la souche B. Parmi les personnes infectées, la souche A est présente à 50 % tandis que parmi les personnes symptomatiques, cette proportion est de 75 %.

Souche	Non-infecté	infecté	symptomatique
Aucune	96 %	0 %	0 %
A	0 %	2 %	1,5 %
B	0 %	2 %	0,5 %

vaccins c'est un peu du blabla".

I : "*En fait, le point que vous soulevez, je n'y avais pas pensé, et je crois qu'on n'en parle même pas dans les documents officiels. De fait, s'il y a une meilleure couverture ou une meilleure efficacité sur... Enfin si la couverture est grande, mais je ne pense pas que la couverture soit très très grande. Mais si tout le monde est vacciné et que, de fait, ça marche très bien pour certaines souches, et très mal pour d'autres alors, de fait, il va y avoir un biais puisqu'on ne va retrouver que les gens (parce qu'on n'analyse que les frottis des gens malades). Mais ici c'est parmi tous ceux qui déclarent la grippe. Tous ceux qui ne l'ont pas déclarée, on ne les analyse pas.*

E3 : "*Donc là, on ne doit pas tenir compte de ce qu'on nous dit sur les vaccins et tout, c'est plus pour nous mettre en situation mais ça ne change rien à la réflexion*".

I : "*C'est ça, on peut simplifier un peu la réflexion, c'est pour ça qu'on dit bon ici il n'y a que deux souches mais en réalité il y en a plus que deux mais on va s'intéresser juste à ces deux-là, c'est comme si on faisait H3N2 versus le reste*".

A travers leurs réflexions, mêmes si elles sont parfois incomplètes ou contradictoires, les étudiants soulèvent la question du biais dans l'évaluation de la proportion des individus infectés par telle ou telle souche. En effet, dans le contexte il est indiqué que la proportion des différentes souches parmi les personnes *symptomatiques* est utilisée pour élaborer le vaccin. Pourtant, on peut soutenir que l'élaboration d'un vaccin devrait se baser sur la proportion de personnes *infectées* par les différentes souches et pas sur les personnes symptomatiques (voir tableau 3.9).

3. **Reformulation du problème par l'investigateur**

   15'30. Les réflexions des étudiants, bien que pertinentes, ne semblent pas de nature à enclencher le raisonnement d'inférence statistique. Pour relancer les étudiants sur ce qui est censé être le cœur de la situation, l'investigateur reformule le problème et présente la machine de hasard.

I : "*(Cette boite) représente les individus malades. Donc quand on fait un prélèvement chez un patient, on va avoir un certain résultat, par exemple ici c'est une boule sans croix donc on va dire que c'est un H1N1 (...) Et donc là-dedans j'ai mis juste 5 balles pour limiter le nombre de combinaisons possibles.(...) Mais la question c'est, avec ce dispositif-là. On a simplifié parce qu'on a dit qu'il n'y avait que deux souches, on a simplifié parce qu'on ne s'intéresse qu'aux gens qui sont déjà atteints sans penser à la vaccination, et on simplifie encore une troisième fois en disant qu'il n'y a que certaines combinaisons possibles.(...) C'est dans ce contexte simplifié qu'on va essayer de résoudre le problème*".

4. **Réflexions générales sur la précision**

E5 : "*Bin alors, on en prend 200 (rires)*".

I : "*Oui, mais il faut justifier parce que vous allez devoir justifier à votre chef que vous demandez 200 prélèvements plus 200 analyses de labos. Il va dire pourquoi 200, pourquoi pas 50 ?*"

E5 : "*C'est comme quand on lance une pièce et que c'est pile ou face, plus on la lance de fois plus on a de chances d'obtenir un résultat moyen qu'un résultat...*"

E4 : "*Ça diminue la marge d'erreur quoi.*"

I : "*C'est comme si on avait des pièces de monnaie, si on veut. (...) Ici on essaie de discriminer six combinaisons possibles (...) on va appeler ça des hypothèses. Dans le cas des pile ou face, c'est comme si on essaie de discriminer une pièce normale d'une pièce truquée. Et un pièce truquée on dirait que neuf fois sur dix elle fait pile. Ok ? Si on a ces deux hypothèses-là, il faudrait voir quel est le nombre de lancers qu'il faut faire pour savoir qu'elle est truquée. Et bien ici c'est pareil, on a six hypothèses, et il faut voir quel nombre de tirages il faut faire pour savoir quelle est la bonne hypothèse*".

5. **Tentative de déduire la proportion de souches H3N2**

E3 : "*Il y a 4 possibilités, il y a 4 virus possibles A H1N1, A H3N2, B Victoria, B Yamagata, du coup ça veut dire que tu as, chaque fois, trois fois plus de chances de pas avoir H3N2 que d'avoir H3N2*"

E3 : "*Ou alors est-ce qu'on doit faire 1/2, 1/2, comme s'il n'y avait que H3N2 et le reste ?*"

E4 : "*Mais c'est pas forcément une chance sur deux puisque les proportions sont pas les mêmes*"

I : "*C'est comme gagner au lotto et ne pas gagner au lotto, c'est deux possibilités mais c'est pas forcément une chance sur deux*".

E5 : "*Oui en fait, donc on ne sait pas prédire*"

E3 : "*Non, justement, la proportion de la chance sur ... c'est ce qu'on veut trouver en fait*"

6. **Approche qualitative de la distribution attendue**

E3 : "*Si t'en prends 10, je me dirai juste que c'est pas possible parce que c'est pas assez, mais après, comment faire la différence entre 200 ou 300, tu vois ?*"

I : "*Pourquoi ($n = 10$) c'est pas assez ?*"

E5 : "*Parce qu'on ne peut pas généraliser une proportion de tous les malades avec un échantillon (aussi faible)*".

E4 : "*On pourrait tomber que sur des gens qui ont l'autre grippe ou soit que sur des gens qui ont le H3N2*".

I : ""*Parce que vous dites, intuitivement, 10 c'est pas assez parce qu'on sent bien que ça peut être aussi bien un résultat qu'un autre. Est-ce qu'il n'y a pas moyen de l'objectiver ? De le mesurer ? On s'attendrait à quel genre de résultats si on faisait 10 tirages ?*"

E3 : "*C'est pas impossible de se retrouver avec que des H1N1 dans les 10 personnes, mais en soi, même avec 100 c'est pas impossible, ça peut toujours arriver qu'on se retrouve que avec des H1N1*".

I : "*Et on ne sait pas le mesurer ?*"

E4 : "*Mais oui mais pour le mesurer il faudrait des valeurs*".

(l'investigateur sort de la pièce)

E4 : "*Tu n'as aucune donnée et le contexte ne te sert à rien*"

E3 : "*En fait clairement (le contexte) on s'en fout. En fait il pourrait ne pas avoir mis tout le bla bla avant et juste avoir dit : combien de prélèvement il faudrait analyser dans des malades pour savoir (. . .) cette maladie-là.*"

E3 : "*Passe un peu la machine. Je pense qu'il faut réfléchir avec ça*".

7. **Tentative de combiner les chiffres de l'énoncé**

E4 : "*On peut avoir une marge d'erreur à 20% près. . . donc 50 (tirages) alors ?*".

E3 : "*50 échantillons ? Pourquoi ?*"

E4 : "*Je ne sais pas, j'ai fait un rapport 5 boules fois 10 (rires)*".

E3 : "*5 fois 20 ça fait 100, ok super (rires)*".

8. **Réalisation de tirages**

29', les étudiants se lancent dans la réalisation de tirages.

E5 : "*On peut faire un test. A partir de quel moment ton échantillon est assez grand pour que ça ne change plus rien ? .*"

E4 : "*En même temps on pourrait faire une expérience là maintenant, avec les boules. Dans le sens que, j'ai réfléchi un peu, s'il y a vraiment 5 boules dedans, on pourrait faire des tests pour les 10, 20, 50. Genre, cela veut dire que tu répètes 10 fois la même expérience, entre guillemets, on fait 20 fois et puis on regarde les marges d'erreurs.*"

E4 : "*Franchement on fait 10, 10, 10, 10, 10 et on verra si c'est toujours pareil*"

E5 : "*Bin déjà là, franchement on voit déjà qu'il y a au moins trois boules sur deux qui sont H3N2*".

E3 : "*T'es passée de 70% à 50%, donc tu as pile 20% de différence. C'est quand même beaucoup passer de 50 à 70%. Du coup, on refait un tirage.*"

E4 : "*Ha mais, non, non, non, on a fait une faute là. C'est pas comme ça qu'il faut regarder les résultats alors (. . .) moi je m'étais dit, genre, t'additionne les deux et tu regarde si tu as le même pourcentage d'H3N2. On ne fait pas par 10, par colonne de 10 (les autres étudiants acquiescent*".

E5 : "*Mais ça fait encore 50%*".

E3 : "*Est-ce que l'échantillon de 10 ce serait assez ?*"

E5 : "*Bin non, parce que ça. . . Ca nous avance à quoi de faire ça ?*"

E3 : (ouvre la bouteille) "*Il y a 3 sur 5 balles où il y a un H3N2*".

E5 : "*Haa tricheuse*".

E3 : "*Mais en même temps, j'arrête pas de lancer le truc, je vois bien. Mais en fait avec le premier lancer ($n = 10$) et le deuxième lancer, on savait dire qu'il y avait plus ou moins 3 balles sur 5 qui avaient H3N2.*"

E4 : "*Là on est à 56% pour 30 lancers*".

E3 : "*Donc c'est quand même mieux d'avoir plusieurs (séries de) lancers. Parce que tu vois ça se stabilise (...) parce que regarde le premier on était à 70%, le deuxième on est à 50% après on est à 50% donc tu vois ça se stabilise autour de 50% alors que le premier était quand même à 70%*".

E4 : "*Donc on peut faire 20 lancers*".

E3 : "*Moi je pense que ce serait 20 ou 50, peut-être 50*".

E5 : "*On en refait 50 alors (on en rajoute 20 pour avoir une fois 50)*".

E3 : "*D'office qu'au début tu veux 200 car ça te paraît grand, mais, de fait, ça coûte cher et tout donc ça ne va pas. 20 et 10 c'est d'office trop petit donc il faut prendre un peu intermédiaire donc moi je dirais 100 ou 50. Et donc 100 c'est quand même beaucoup et 50, voilà (c'est le nombre qu'il faudrait choisir).*"

E5 : "*60% (4e série de 10 lancers).*"

E3 : "*Donc en fait voilà, c'est 50 parce que 10 c'est clairement pas assez, on voit que la différence (entre les échantillons de taille 10) est trop grande. 20 c'est bien mais vu qu'on s'est arrêtés à 20, on ne peut pas être sûrs que c'est à partir de là que ça se stabilise alors que après 50 on voit que là clairement il y a une stabilisation*".

E4 : "*Oui mais on voit après 30 (lancers) qu'il y a une stabilisation*"

E3 : "*A 20, tu as atteint le juste nombre de ta proportion*".

E5 : "*Non non c'est à 30 que tu (l') as atteint.*"

E3 : "*A 20, on était déjà à 50 et des...*".

E5 : "*Oui mais c'est parce qu'on a eu de la chance, il y a un facteur chance*".

E3 : "*A 20, tu peux atteindre la bonne proportion mais tant que tu n'as pas fait de lancers supplémentaires, tu ne peux pas voir que c'est cette proportion-là qui est stabilisée. Du coup, t'es obligée d'aller quand même plus haut même si t'as le bon nombre pour être sûre qu'à tes tests d'après ce soit cette proportion qui se stabilise.*"

TABLE 3.10 – **Résultats des tirages du groupe 2**. F.R. = Fréquence relative, F.R.C.= fréquence relative cumulée.

Tirage	F.R.	F.R.C.	Différence de F.R.C.
1	5/10	50,0 %	
2	7/10	60,0 %	10,0 %
3	5/10	56,7 %	3,3 %
4	6/10	55,0 %	1,7 %
5	6/10	56,0 %	1,0 %

47', ils récapitulent leur raisonnement devant l'investigateur.

I : "*Mais ici vous l'avez fait une fois ($n = 50$). Cette fois-là ça s'est stabilisé mais peut-être que si on le refaisait, il faudrait 100 lancers pour que ça se stabilise, non ?*"

E5 : "*Non parce que franchement ça se voit que c'est stable*"

50'05. Hormis, par moment l'étudiante E5 qui soulève parfois la question du *facteur chance*, les étudiants semblent satisfaits de leur raisonnement. Ils ont une assez bonne confiance dans le fait qu'une répétition de l'expérience ne mènerait pas à des variations très différentes de ce qui a été observé. Le but était d'avoir une marge d'erreur de moins de 20 % et ici les estimations faites entre $n = 20$ et $n = 50$ ne varient que de quelques pour cents (voir tableau 3.10).

9. **De la variabilité intra-séries vers la variabilité inter-séries**

   En somme, les étudiants décrivent ici une variabilité à l'intérieur d'une série de 50 lancers, une variabilité intra-série. L'investigateur les invite maintenant à considérer la variabilité inter-séries.

   I : "*C'est bien (...) intuitivement c'est assez correct (...) mais il manque de quantification pour dire : voilà pourquoi on choisit 40 ou 60 ou 80. (...) Ici vous avez une série de 50 lancers*".

   E3 : "*Mais il faudrait montrer que sur chaque série de 50 lancers ça fonctionne*".

   I : "*Ou sur plein de séries de 50 lancers, oui*".

   E4 : "*Donc on pourrait faire une dizaine, c'est-à-dire 500 tests là*".

   E3 : "*Ou alors il faut mettre un truc pi machin là*".

   A partir de ce moment-ci (52'00), l'investigateur va amener, de manière plus directe, les étudiants vers un des raisonnements attendus, passant par les tables de la distribution binomiale.

10. **Proposition d'utilisation de la distribution binomiale**

    I : "*S'il y avait effectivement 3/5 de H3N2 et qu'on faisait 50 lancers, quels seraient les résultats qui seraient probables et quels seraient les résultats qui seraient improbables ?*".

    E3 : "*Et bien par exemple d'avoir 1 H3N2*".

    I : "*Oui, de nouveau je suis d'accord, mais il y a une manière de les quantifier. Parce que, bon, pour 1 on va tous être d'accord pour dire que 1/50 (résultat observé) si c'est 3/5 (proportion vraie), ça n'arrivera jamais.*"

    E5 : "*Ho, il faut faire un arbre.*"

    E3 : "*(Il faut calculer la) probabilité que $X = 0$, si on dit que X est la chance entre guillemets d'avoir H3N2, la probabilité que $X = 0$ est proche de 0*".

    I : "*Oui c'est ça. Mais il y a moyen de les trouver (ces probabilités). Allez, je vous aide (sort un formulaire contenant notamment les tables binomiales)*".

11. **Utilisation des formules puis des tables liées à la distribution binomiale**

    Les étudiants essaient d'utiliser les formules de la distribution binomiale mais le fait de poser une hypothèse en plaçant π à 60 % semble contre-intuitif. D'où vient ce nombre ?

E3 : "*En fait on a déterminé nous que c'était 60 %. Mais ça on aurait dû savoir le faire déjà avec la formule ? Parce que là, par exemple, on pourrait mettre 'espérance d'avoir 60 %*".

12. **Construction d'intervalles de décision excluant les valeurs rares sous une hypothèse**

01h05, l'investigateur rappelle le fonctionnement de la table binomiale puis amène l'idée d'intervalles de décision.

I : "*On a fait ici une série de 50 tirages. Et donc on a obtenu cette fois-là, on est tombés très près de 60%. Et la question c'est, si on faisait plein de séries de 50 tirages, combien de fois on serait très proche des 60% et combien de fois on s'en éloignerait ?*"

E4 : "*Il y a des fois on serait peut-être éloignés, il y a des chances que l'on soit éloignés. . .*"

I : "*Et ça de nouveau on ne peut pas le mesurer ?*"

Pendant une dizaine de minutes, les étudiants cherchent à déterminer la zone des résultats probables si $\pi = 0.4$ et $n = 50$.

E5 : "*Bin, c'est ce que j'avais dit, entre 15 et 25.*"

E4 : "*Donc on a une brochette, et maintenant qu'est-ce qu'on fait avec ça ?*"

E5 : "*Donc on a une marge de 20% (avec $n = 50$, car résultats possibles tombent entre $x = 15$ et $x = 25$ soit entre 30 et 50%)*".

I : "*Comment tu as choisi 3 et 15 (zone acceptation $\pi = 20\%$ avec $n = 50$) ? (l'étudiante cherche), Comment on va choisir ces valeurs-là ?*"

E5 : "*Bin avec les probabilités, (. . .) c'est quasi improbable d'avoir un 3 ($n = 20$, '$\pi = 40$ %)*"

I : "*Et tu mets où ta limite de quasi-improbable ?*"

E5 : "*Bin plus ou moins à 1 %*".

E5 : "*(du coup les limites pour $n = 20$, $\pi = 40$ % seraient) plus ou moins 4 et 12 alors (. . .) et donc là ça ferait entre 20 % et 60 %, et du coup ça ferait une marge d'erreur de 40 % et là ($n = 50$) une marge d'erreur de 20 % et nous on cherche la marge d'erreur de 20 % donc voilà (exercice terminé).*"

13. **Fin de l'exercice**

1h33, discussion sur la fin de l'exercice.

E5 : "*On doit encore faire quelque chose ou on a trouvé ?*"

I : "*(je n'ai pas l'air convaincu) car je ne sais pas si vous êtes vous mêmes convaincus*".

E3 : "*Ha si,mais moi déjà avec notre expérience des lancers, j'étais déjà convaincue.*"

E3 : "*Je ne sais pas, les tables ça me paraît fort abstrait. Enfin pas abstrait mais (. . .) j'ai l'impression que ce n'est pas vrai*".

E3 : "*Heureusement quand même qu'on avait l'expérience avant sinon on n'aurait peut-être pas compris à quoi correspondaient les tables. Si on nous avait mis direct(ement) devant les tables, on n'aurait pas compris je crois*".

14. **Résumé**

01h45, une des étudiantes résume la démarche.

E3 (pour le groupe) : "*Donc, au début, on a d'abord pensé à prendre 200 parce qu'on s'est dit qu'au plus l'échantillon est grand au plus on est précis sauf qu'il y a un coût quand même, des obligations et tout ça. Donc, on s'est dit qu'on pouvait peut-être prendre plus petit puisqu'on pouvait avoir une marge d'erreur de quand même 20 %. Du coup, on a fait une expérience avec au début 10 lancers, puis 20 lancers puis 50 lancers. L'expérience nous a montré qu'après 20 lancers, la différence avec les 10 premiers était de 20 % mais c'était quand même élevé donc on s'est dit qu'on allait continuer. Puis au dessus de 20 lancers, donc une fois qu'on est arrivés à 30, 40 on a vu que la proportion se stabilisait. Donc on a conclu qu'il fallait à peu près 50 lancers, dans les propositions données dans l'énoncé. Puis après, on a complété notre raisonnement par les tables binomiales où on a vérifié notre raisonnement en comparant les marges d'erreurs quand* $N = 20$ *ou* $N = 50$. *Et on a vu que du coup on avait 95 % de chances d'avoir entre 12 et 26/50 de non-malades. Notre marge d'erreur c'était 28 % du coup c'était quand même élevé mais vu qu'on avait quand même 95 % de chances et bien, ça allait. Et entre 3 et 14/20 de non-malades donc 40 % et du coup ça confirmant nos plus ou moins 60 % qu'on avait dans notre expérience avant avec nos 50 lancers*".

01h48, fin de la séance.

Groupe 3

1. **Recherche d'une piste par analogie**

   Dans un premier temps (durant 25'), les deux étudiantes essaient de glaner des informations afin de savoir quelle piste prendre. Elles cherchent à rapprocher ce problème des exercices qu'elles auraient déjà rencontrés auparavant ou bien à identifier dans la suite du syllabus (dans la matière qui n'a pas encore été vue) des éléments de réponse.

   E6 : *"Je crois qu'on va devoir attendre qu'il nous dise par où commencer"*

   I : *"Vous avez compris ce qu'on cherche ?"*

   E6 : *"Combien d'échantillons il faudrait prélever pour avoir la probabilité de trouver une fois une partie du virus A H3N2. Mais du coup, je me suis dit qu'on pourrait peut-être utiliser le théorème limite central mais il manque plein de trucs pour."*

   E6 : *"on dit que pour le vaccin, elle (l'efficacité) est faible ou nulle mais qu'est ce que cela veut dire faible ou nulle ?"*

   I : *"Ici on va juste s'intéresser aux gens qui sont déjà malades, pas à tous ceux qui ont échappé à la maladie grâce à leur vaccin. On va juste prendre, dans tous ceux qui sont malades, il y a quelle proportion de H3N2, quelle proportion de H1N1, et on va laisser de côté les autres sous-types."*

   E7 : *"Donc on doit commencer par faire des tirages aléatoires ?"*

   I : *"On peut. Par exemple, ici, on va en faire un, voilà (. . .) "*

   I : *"On va travailler dans ces conditions-là,*

c'est-à-dire qu'il n'y a que deux possibilités : H3N2 ou H1N1, et la composition de la population (boite) ne peut être que... vu qu'il n'y a que 5 balles dedans, il n'y a que certaines compositions possibles. En gros, il ne peut pas y avoir 13 % de H3N2 vu que je n'ai mis que 5 balles. (...) Donc on a simplifié un peu le problème pour que ce soit plus facile de démarrer."

I : *"Ça (la composition) peut être 0, 20, 40, 60, 80 ou 100 %. Donc on a simplifié un peu le problème pour que ce soit plus facile de démarrer".*

I : *"Donc on a tout ça comme possibilités (les compositions possibles) et la question c'est 'il faudrait faire combien de tirages pour pouvoir les... savoir laquelle est bonne ?'".*

E6 : *"J'ai compris le but mais je n'ai pas du tout compris comment on allait faire ça... "*

E7 : *"C'est par rapport à quelque chose qu'on a vu en cours du coup ? Ou on ne l'a pas encore vu ?"*

E6 : *"Est ce que ça servirait à quelque chose de faire toutes les combinaisons possibles et du coup de voir les probabilités de chacun ? (pourquoi pas) C'est ce qu'on a fait au cours passé."*

E7 : *"Moi je ne vois pas du tout comment on démarre."*

E6 : *"J'ai l'impression qu'on n'a rien comme information."*

E6 : *"Je n'ai jamais rien compris avec ces matières-là".*

E6 : *"Ça me paraît bien dans le sens où il faut déterminer N, c'est un des seul trucs que je connais où il y a N."*

E6 : *"Mais je me souviens qu'il fallait faire autre chose... "*

E7 : *"Mais attend, ça date, la rhéto c'est loin."*

E6 : *"C'est (perturbant) d'avoir des pour cent là (dans l'énoncé) et de ne pas les utiliser"*

E7 : *"C'est peut-être juste pour l'intro."*

E7 : *"Regarde un peu dans les modules qu'on n'a pas encore faits."*

E7 : *"Tu crois qu'il va nous taper un truc comme ça à l'exam ?"*

E6 : *"Ho, moi ça me saoule, je n'aime pas être bloquée comme ça".*

E6 : *"La probabilité d'avoir l'un ou l'autre c'est 50 %."*

E6 : *"Et si on faisait avec le truc carré là, chi-carré ?"*

2. **Raisonnement hypothético-déductif suggéré par l'investigateur**

27', cette première phase, au cours de laquelle l'investigateur a tenté de ne pas intervenir dans le choix de la piste à emprunter pour résoudre le problème, conduit à un blocage et à une certaine démotivation. L'investigateur intervient de manière plus claire et tente de les placer sur la piste d'un raisonnement hypothético-déductif.

I : *"Si on prenait 10 tirages, et que la proportion c'était, mettons, 1/5e, on aurait quel genre de résultats attendus ? Qu'on pourrait observer s'il y a 1/5e des gens qui ont le H3N1 dans la population, dans l'échantillon, sur les 10 on pourrait en observer 0, ou 1 ou 2 ou 3 ... et ainsi de suite jusque 10."*

I : *"On pourrait en observer 0 tout comme on pourrait en observer 10 (H3N2 en 10 tirages avec $\pi = 20\,\%$). Le plus probable ce serait peut-être 2, non ?. Pas loin de 2, ou entre 0 et 4. Et au-dessus de 5 ce serait peut-être difficile d'obtenir 5 personnes qui ont le H3N2 sur 10 si dans la population il n'y avait qu'une seule balle avec le H3N2, d'accord ?".*

E7 : *Oui (perplexe)*

E6 : *"Comment est-ce qu'on peut calculer la probabilité (pi) ? C'est ça qui me perturbe. Parce qu'on n'a aucun pourcentage."*

E6 : *"Est-ce qu'on peut déjà partir du principe qu'on a un chance sur deux, non ?"*

E6 : *"Et puis même si on fait 10 tirages, est-ce qu'on peut considérer que c'est vraiment la bonne probabilité ou pas ? Parce qu'on pourrait obtenir n'importe quoi."*

I : *"Et si on fait 10 tirages, on pourrait obtenir quoi ?"*

E6 : *"On pourrait obtenir très bien 10 balles avec une croix ou on pourrait en avoir aucune. Donc en faisant des tirages comme ça on peut... si chacun fait différemment on peut avoir des probabilités différentes. Du coup je ne comprends pas comment on peut avoir une probabilité juste."*

E6 : *"De base on était partis avec la loi binomiale mais on ne s'en sort pas trop."*

L'investigateur rappelle l'utilisation de la table binomiale

I : *"Je vais vous dire comment on lit la table binomiale. (...) en fait il y a deux 'si' pour la table binomiale. Si on prenait 10 tirages (...) et si la proportion dans la population est de 1/5."*

E7 : *"Et ça on définit comment ?"*

I : *"Bin, ici il faut dire 'si c'est ça la proportion', voilà les résultats attendus"*

E7 : *"Ok"*

I : *"S'il y avait une balle sur 5 qui était H3N2, alors en faisant 10 tirages on tomberait parfois sur 0 (H3N2/10) puisque la probabilité serait de 10 % (...) donc ça veut dire que sur 10 lancers, il y a 10 % de chances d'obtenir un 0, et il y a 37 % de chances d'obtenir 0 ou d'obtenir 1..."*

E7 : *"Donc du coup, on pourrait très bien faire 10 tirages avec la balle, prendre la probabilité qu'on a et dire que c'est celle-là ? (parce que) de nouveau n'importe qui pourrait obtenir autre chose du coup, ça me perturbe."*

E6 : *"Du coup, c'est mieux de prendre la table où il y a le plus de lancers (oui pourquoi ?) parce que ce sera plus précis (qu'est-ce qui sera plus précis ?) la valeur du nombre de personnes atteintes par A H3N2 (l'estimation)."*

I : *"Si vous voulez, on peut faire des tirages."*

E6 : *"Mais ça restera toujours le même problème qu'on peut obtenir n'importe quoi. Mais comment est-ce qu'on pourrait régler ce problème-là ?"*

I : *(Si $\pi = 20$ % et qu'on fait 10 tirages, $P(X = 0) = 10$ %).*

E6 : *"Mais comment est-ce qu'on peut savoir qu'on a 20 % de chances de l'avoir ? Comment est-ce qu'on sait que la probabilité c'est 0,2 ?"*

I : *"Mais ici j'ai dit 'si c'était ça'."*

E6 : *"Mais oui mais justement il faut surement savoir lequel. Mais comment le calculer ou le déterminer ?"*

3. **Introduction du principe de réfutation d'une hypothèse par l'observation de résultats rares**

Après 39', l'investigateur introduit l'idée que des résultats rares sous une hypothèse la contredisent, permettent de l'exclure.

I : *(ici il faut savoir comment discriminer les 6 possibilités ($\pi = 0$, 20 %...). Parmi les valeurs possibles, il faut voir s'il n'y a pas des résultats qui permettent d'exclure certaines hypothèses et de garder les autres... Je vais vous donner un exemple : avoir 0 H3N2 en 10 tirages, sous l'hypothèse que c'est 1/5e de H3N2, il y a 10 % de chances que ça arrive, autrement dit ça peut arriver. Mais avoir 0 H3N2 en 20 tirages, là il y a 1 % de chances que ça arrive. Du coup si on prend 20 tirages et qu'on a 0 H3N2 sur les 20, ça peut difficilement arriver sous l'hypothèse qu'il y a 20 % de H3N2, ça peut encore moins arriver sous l'hypothèse qu'il y a 40 % de H3N2 et encore moins pour les autres (hypothèses), par contre ça peut arriver sous l'hypothèse qu'il y a 0 % de H3N2. Du coup, le résultat n'est pas compatible avec la proportion 20 %, (ni 40, 60 80 %), il n'est compatible qu'avec 0 %."*

E6 : *"Et donc en faisant vraiment les tirages..."*

I : *"Si on avait 0, sur 20 tirages, on pourrait dire (que) c'est certainement (cette hypothèse-là qui est correcte). On pourrait éliminer toutes les autres hypothèses sauf l'hypothèse qui dit qu'il y a 0/5 H3N2."*

E6 : *"Moi je ne suis pas sûre d'avoir compris"*

I : *(Le résultat $X = 1$ ($N = 20$) est compatible avec $\pi = 20$ % mais pas avec 0 ni 40 %)*

E6 : *"Mais ça va nous mener à quoi ça ?"*

E7 : *"Du coup, si on se rend compte, dans la table $N = 20$, que c'est 0,2 la plus probable on peut partir du principe que pour $N = 10$ c'est 0,2 aussi ?"*

I : *"Ha pour savoir quelle composition il faut choisir et bien il faudra faire un tirage à un moment..."*

E6 : *"Moi j'ai compris son raisonnement mais je n'ai pas compris ce qu'on peut faire avec"*

4. **Réalisation de tirages et interprétation**

53', les étudiantes réalisent des tirages. La fréquence de H3N2 est de 12/20, puis 14/25, puis 3/5, donc 29/50 au total.

I : *Etait-ce possible sous $\pi = 20$ % ?*

E7 : *"Bin donc 0,4 (sur base de ce résultat), c'est pas possible. C'est d'office au dessus. Comment maintenant on va savoir si c'est 0,6 ou 0,8 ?"*

I : *"En fait pensez que 60 % de H3N2 c'est la même chose que 40 % de H1N1."*

(Les étudiantes éliminent $\pi = 1$ sur base d'un raisonnement logique).

E6 : *"3 chances sur 5 de tomber sur le virus H3N2. (ok comment l'a-t-on déterminé ?) En regardant la probabilité de tomber sur plus ou moins que 3 chances sur 5. (vraiment ?) On a regardé que, par exemple 0,4 la probabilité était pas très grande et que 0,8 non plus, du coup 0,6".*

I : *"Et du coup comment pourrait-on dire si $N = 20$ est suffisant ou $N = 50$?"*

E7 : *"Est-ce que ce n'est pas, bêtement, parce que quand on avait 10, 20 et 25 l'imprécision était trop grande et que du coup on ne savait pas déterminer lequel était bon ? Tandis qu'avec 50, il y a plus de précision donc ça va."*

I : *"Et quand tu dis 'l'imprécision est trop grande', ça veut dire quoi ?"*

E7 : *(Utilise la $P(X = x)$ pour déterminer si une hypothèse est vraisemblable)*

5. **Détermination d'intervalles de décision**

01h13', l'investigateur amène l'idée d'utiliser des intervalles de décision.

I : *"Imaginons qu'on prenne 25 tirages, quels sont tous les résultats qu'on pourrait observer si la probabilité était de 0,2" ?*

E6 : *"De 3 à 25 ?"*

I : *"Et pourquoi tu exclus 2 ?"*

E6 : *"Parce que c'est trop peu"*

I : *"C'est quand même 10 %, non ?"*

E6 : *"Mais ça dépend à partir d'où on définit que c'est trop peu"*

I : *"Donc si on fait 25 tirages, des résultats plausibles sous cette hypothèse-là c'est tous ceux-là (x est contenu dans [2-8]). Et si c'était l'hypothèse de 2/5, ce serait quoi les résultats plausibles ?"*

E6 : *"A partir de 6 jusque 14"*

E7 : *"Parce que tu n'as pas plus de 5 (%) entre les deux ($X = 13$ et $X = 14$?)."*

I : *"(attention) c'est pas la différence qu'il faut regarder ($P(X = x)$) c'est combien il reste ($P(X \geq x)$)."*

I : *(Avec $N = 20$, si $\pi = 20$ %, les résultats plausibles vont de 2 à 8, si $\pi = 40$ % ils vont de 6 à 14, du coup il y a des résultats qui (ex 7) qui sont compatibles avec les deux hypothèses)*

E6 : *"Et du coup dans 50 ça ne se recouperait pas du tout ? Ce serait deux intervalles complètement différents comme ça on sait d'office les différencier ?"*

I : *"Essayons."*

(Les étudiantes déterminent les bornes entre 6 et 15 ($\pi = 20$ %) et 14 et 25 ($\pi = 40$ %)).

E7 : *"Donc si tu tombes sur 14 ou 15, c'est ambigu aussi"*

I : *"Et vous avez quelle probabilité de tomber dans votre intervalle ?"*

E7 : *"On a 92 % ($\pi = 0,4$) de tomber dans l'intervalle. (et pour $\pi = 0,2$), on a 87".*

E6 : *"C'est trop proche"*

E7 : *"Du coup ça ne va pas du tout non plus ?"*

E6 : *"Du coup il faut plus, mais on n'a pas des tables de plus".*

6. **Relation entre la taille d'échantillon et la probabilité d'erreur**

01h21', L'investigateur introduit l'idée d'une relation entre le nombre de tirages et le risque d'erreur.

I : *"Mais il y a deux manières de résoudre le problème, soit on monte encore un petit peu, on va aller jusque 100 par exemple et voir si à 100 on a bien des zones qui ne se recouvrent pas, oui ? Soit on augmente un petit peu la tolérance."*

E7 : *"Ha, on pourrait monter à 10 % par exemple."*

I : *(propose de monter l'erreur alpha pour que ce soit bon avec $N = 50$)*

E7 : *"Mais ce que je ne comprends pas c'est que, si on fait pareil avec la table de 25, qu'on réduit la tolérance par exemple à 20 % bien ça va peut-être marcher aussi du coup ? (tout à fait) Un moment si on réduit la tolérance à 50 % ça ne va pas nous avancer à grand chose, si ?".*

I : *"Tu as raison, c'est juste qu'il n'y a pas une valeur (alpha) qui est plus justifiée qu'une autre. Si on choisit telle, ce qu'on va appeler, "tolérance", alors c'est telle taille d'échantillon qu'il faut."*

E7 : *"Donc ça pourrait être n'importe lequel en fonction du modèle qu'on prend. (oui) Mais du coup toutes les réponses sont justes ? (oui) Haaaa (je comprends)."*

7. **Reformulations et discussion à propos du principe utilisé**

   01h23, par moments les étudiantes semblent suivre le raisonnement proposé par l'investigateur mais, par moments, l'étudiante E7 semble penser que l'on peut déterminer le contenu de la boîte uniquement en regardant les tables, sans réaliser le moindre tirage.

   E6 : *"Donc on ne sait pas répondre à la question."*

   I : *"Si mais vous pouvez dire 'si on accepte ça comme risque d'erreur' alors ce serait telle taille d'échantillon la bonne. Si on veut moins de risque d'erreur alors il faudra sans doute augmenter la taille de l'échantillon".*

   E7 : *"Du coup, moi ce qui me perturbe c'est que pour la question on peut répondre n'importe quelle valeur du moment qu'on justifie correctement."*

8. **Raisonnement alternatif à partir d'intervalles fixes**

   01h25', l'investigateur propose un autre raisonnement qui utilise des intervalles de décision fixes et une probabilité d'erreur qui dépend du nombre de tirages.

   E7 : *"Mais je ne comprends pas pourquoi on fait ça"*

   E6 : *"Juste pour tester une autre technique."*

   E7 : *"Oui mais j'aimerais bien comprendre à quoi ça sert."*

   E7 : *"On place les limites des intervalles pour être sûrs qu'ils ne se recouvrent pas et puis on regarde quelle est la probabilité de tomber dedans."*

   E7 : *"Et du coup si on a une grande probabilité d'être entre 5 et 15, ce serait 20 %".*

   I : *"Non, on pourra dire que c'est 0,2 si notre résultat tombe quelque part là-dedans. Si on fait les tirages et qu'on a 6 H3N2 alors on conclura que c'est celle-là qui est la bonne."*

   E7 : *"Ok. Mais là on va quand même regarder laquelle a la plus grande probabilité, non ? (heu, développe un peu). Mais quand on a défini ça, il fallait de nouveau regarder l'écart entre les deux pour définir la probabilité ? (oui) Et du coup c'est en fonction de laquelle est la plus grande qu'on peut définir quel modèle est le bon ?".*

   I : *(Réexplique l'idée de calculer erreur à partir d'intervalles fixes).*

   E7 : *"Et en ayant les probabilités, on va faire quoi après avec ça ?"*

   E : *(ne s'y retrouvent pas avec le raisonnement utilisant des intervalles fixes, repartent sur la piste d'augmenter l'erreur tolérée en fixant la taille d'échantillon).*

9. **Fin de l'exercice**

   E : *(Fixent alpha à 20 %, constatent que* $N = 50$ *est suffisant)*

   E7 : *"Mais sinon on a fini le problème, là, du coup ?"*

10. **Résumé**

    01h48, les étudiantes résument la démarche.

"Donc on a fait les tirages pour $N = 20$, puis pour $N = 25$ puis pour $N = 50$. Du coup, après on a regardé dans les tables binomiales pour $N = 20$ d'abord, et on voyait que c'était semblable entre le 1 chance sur 5 et le 2 chances sur 5. Donc on ne savait pas déterminer à partir de ça. Et du coup on a pris le $N = 25$ et on a vu que le 1 chance sur 5 n'était pas possible, parce que ça tombait sur 1, je ne sais plus... Et du coup, c'était 0,4 ou plus. Et donc dans le N 50, on a vu que 0,4 n'était pas possible, on devait voir si c'était 0,6 ou 0,8. Donc on a fait en inversant, en regardant pour l'autre virus en faisant 1-0,8 donc en regardant à 0,2. Et c'était pas possible donc c'était d'office 0,4 pour l'autre virus et 0,6 pour le virus qu'on recherchait. Et donc en conclusion, on avait 3 chance sur 5 de tomber sur le virus H3N2. Et pour répondre à la question pour savoir si l'échantillon est suffisant, on choisit l'intervalle en laissant par exemple 5 % de risque d'erreur au dessus et en dessous, je ne sais pas comment on appelle ça, des deux cotés, en regardant pour 0,2 et 0,4 et on voit que les intervalles se regroupent. Du coup, soit on augmente N, soit on augmente le risque d'erreur pour avoir des intervalles qui ne se regroupent pas. Et donc la réponse à la question c'est par exemple : $N = 50$ est suffisant si on tolère 20 % de risque d'erreur en tout."

Groupe 4

1. **Clarifications du contexte**

   Assez rapidement, l'investigateur tente de lever des difficultés qui pourraient être liés au contexte pour lancer les étudiantes dans ce qu'il considère constituer le cœur du problème.

   E8 :*"Au final, le fait que ce soit des sortes de virus, limite on s'en fout."*

   I :*"Vous voyez un peu (comment démarrer) ?"*

   E9 : *"Moi je ne sais pas très bien comment partir en fait"*

   E8 :*"Ici... sur le tirage aléatoire de cinq sous-types (quid ?)"*

   I :*"Ça c'est en rapport avec la petite boite. L'idée c'est qu'on va simplifier le problème. Plutôt que de se dire qu'il y a quatre sous-types différents et toutes de proportions possibles différentes, on va travailler dans un modèle simplifié où il n'y aurait que deux sous-types : les H3N2 et les autres."*

   E8 :*"Ha ok"*

   I :*"Alors ici là dedans j'en ai mis cinq au hasard, j'ai mis des H3N2 et des H1N1, mais il n'y en a que cinq, d'accord ? Donc ceci ça va représenter la population et il y a une certaine proportion de H3N2 dedans qu'on ne connaît pas tout comme dans la population des gens qui sont atteints de syndromes grippaux il y a une proportion de H3N2 qu'on ne connait pas (oui). Ceci sert à représenter la population et on peut faire des tirages pour voir si on est en présence d'un prélèvement où il y a du H3N2, symbolisé ici avec une croix, ou pas. Vu qu'il n'y a que cinq balles ça limite les proportions possibles dans la population".*

   E8 :*"Parce qu'en fait, ici on nous demande combien de prélèvements faudrait-il analyser, c'est par exemple combien de balles on devrait tirer pour extrapoler au final ?. Donc ici maximum cinq et minimum 1"*

   I :*"Sauf qu'on les remet dedans (ha ok)"*

   E8 :*"Bin du coup, si on les remet de-*

dans, on pourrait retirer une autre fois la même et donc ça augmenterait les chances qu'on... Imaginons, donc il y en a cinq dedans, il y en a trois H3N2 et deux H1N1. Et je prends cinq fois la H3N2, ça va faire comme s'il y en avait cinq dedans. Enfin je ne sais pas j'ai un peu du mal avec ça... "

I :*"Et bien, en fait, il faut se dire que ce truc-là va donner des échantillons, puisqu'on fait des tirages cela va donner des résultats parcellaires, on appelle ça des échantillons, avec des proportions qui sont variables. Tout comme si on allait dans la population et qu'on prenait X patients, ça nous ferait un échantillon avec une proportion qui est variable, selon l'échantillon qu'on va prendre on ne va pas avoir les mêmes proportions (ok). Et derrière la proportion de l'échantillon on va essayer de savoir quelle est la proportion dans la population. (Oui, bin c'est ce qu'on nous demande). Tout comme ici derrière les fréquences qu'on va obtenir, on va essayer de savoir ce qu'il y a dedans. (ok)"*

I :*"On a essayé de simplifier le problème général pour que ce soit plus facile à traiter".*

2. **Tentative de résolution du problème par les formules d'analyse combinatoire** 10', dans un premier temps, les deux étudiantes cherchent les formules qui leur permettraient de résoudre le problème dans le champ de l'analyse combinatoire.

E9 : *"Comment ça s'appelait ce qu'on a fait en rhéto ? (des probabilités ?) non l'autre, l'analyse combinatoire (heu ouais, ça me dit quelque chose mais...)."*

E9 : *"Tu ne te souviens pas qu'on faisait des trucs genre avec ... et on prenait un nombre ici... et c'était vraiment choisir parmi ça et ... tout prenait en compte... et selon la lettre, si tu remettais ou pas... Moi je suis sûre que c'est de l'analyse combinatoire (si si)".*

E8 :*"Moi j'ai quand même une question, je suis sûre qu'on a vu dans le même genre en rhéto, l'analyse combinatoire. On devait choisir, justement il y avait plusieurs lettres et il fallait juste rentrer les variables et on trouvait... Mais bon à l'époque j'avais compris, je comprenais tout et j'avais toutes les formules devant moi et là ce n'est pas du tout le cas, et je ne sors pas de rhéto il y a six mois donc heu... "*

I :*"Non, j'imagine, mais il y a quelques formules qui sont là-dedans si vous voulez (fournit le formulaire du cours de statistique qui contient notamment les formules de la loi binomiale)."*

E8 :*"Mais en fait il faudrait juste que j'aie mon cours et je vous trouve le problème".*

I :*"Si vous voulez aller sur Internet, vous pouvez, je vais vous ouvrir une session invité"*

E9 : *"Mais on part sur quelque chose de pas mal ?"*

I :*"Il y a plein de manières d'aborder le problème (ha ok [ok il ne nous en dit pas plus])."*

En cherchant sur internet dans des exercices d'analyse combinatoire, les étudiantes tentent de trouver comment appliquer les formules vues au problème qui leur est posé.

E9 : *"Mais, en gros, ici il faudrait quand même que l'on sache ce que représentent tous ces X, N et tout dans la formule. Moi j'ai l'impression que c'est ça, c'est une formule de ce genre.*

(mais oui oui, mais moi je pense qu'on va y arriver). T'avais ton ensemble, ton nombre de variables possibles . . . "

E9 : *"Oui mais il nous manque deux trucs (N nécessaire, π ?). (. . .) Moi j'ai pas l'impression que ça s'applique super fort. Ou peut-être qu'on doit passer par plusieurs. . . "*

E8 :*"Et il n'y a pas beaucoup de données hein quand tu regardes".*

3. **Clarifications relatives à la machine de hasard**

26', un des points qui pose problème aux étudiantes est la machine de hasard. L'investigateur tente d'expliquer ce qu'elle représente.

E8 :*"Moi ce que je ne comprends pas, c'est la population, j'arrive pas à me dire qu'on travaille sur cinq sous-types. Enfin j'ai pas compris le . . . (. . .) Les cinq, c'est les personnes ou pas ?""*

I :*"En gros, on a fait une population, on va considérer que ce truc-là, c'est une population puisqu'il peut en sortir un nombre de valeurs aléatoires."*

E8 :*"C'est la boite la population ? (oui) Ha ok, mais on ne sait pas le nombre de personnes qui ont été. . . "*

I :*"Tout ce que l'on sait, c'est que, vu comment est construite la boite, les proportions de H3N2 (viennent une fois sur cinq ?), et bien elles sont réduites, cela ne peut que être de 0/5, . . . (1, 2, 3, 4 ou 5 fois, ok)."*

E8 :*"Et avec une formule, ou avec . . . on sait déterminer ça ou il faut expérimenter et voir ? (. . .) Pour connaître la proportion, est-ce qu'on devrait jouer avec la boîte et regarder ou on peut déterminer ça par calcul ?"*

4. **Discussion autour de la précision**

28' Les étudiantes donnent une réponse intuitive et pertinente basée sur l'idée de précision.

E9 : *"En soi, au plus on en fait (de tirages) au plus on sera précis (oui). Donc on peut en faire un et avoir une certaine proportion et en faire 50 et avoir une proportion plus précise".*

E8 :*"C'est ça que le 10 et le 20 on peut les. . . (exclure)"*

I :*"Il faut savoir où l'on s'arrête, est-ce qu'on s'arrête à 200, à 2000, à 20.000 ?"*

E9 : *"Bin si on veut être très, très, très précis on fait toute la population"*

E8 :*"Mais on ne va pas faire ça ?"*

E9 : *"Combien de prélèvements faudrait-il analyser pour connaître la proportion dans la population ? Bin tous si possible, un maximum aussi si possible".*

I :*"Tu as raison, c'est correct. Après on se rend compte effectivement qu'il va y avoir un compromis à trouver parce qu'on ne pourra pas les prendre tous et donc il faudra qu'on. . . Enfin comme tu as dit, au plus on va prendre de personnes au plus ce sera précis. Donc la question ça va être. . . . quelle précision il nous faut ?"*

E9 : *"Bin ça vous nous le demandez, à vous de nous le dire ? Parce qu'en soit, quelle précision il nous faut, bin la plus précise possible. . . ".*

I :*"Et bien regarde quelles sont les compositions possibles dans cette population-ci ? Toutes les compositions ne sont pas possibles puisqu'on a dit qu'on a construit la machine d'une certaine manière (comment ça toutes les compo-*

sitions ?). Bin ça ne peut que être...

E9 : *"H1 avec N1 et pas H1 avec N2, ou c'est pas ça que vous voulez dire ?"*

I :*"Non, je pensais que ça peut être 20 % de H3N2 ou 40 % mais ça ne sera pas 35 par exemple, vu qu'il n'y a que cinq balles. (ha oui, mais je ne comprends pas en quoi ça.[nous aide]).Et bien on a une hypothèse à 20 %, une hypothèse à 40 %, une hypothèse à 60 %, ... Qu'est-ce qui nous faudrait comme précision pour savoir ce qu'il y a dans la boite ?"*

E8 :*"Je suis désolée mais je ne comprends pas le... "*

5. **Reformulation de la question en termes de choix**

31', l'investigateur tente de faire le lien entre la précision nécessaire et le besoin de choisir une hypothèse à l'issue de l'expérience.

I :*"Donc les hypothèses à départager, c'est (...) 0/5, 1/5, etc. Ok donc les hypothèses se séparent de 20 %. Vu que l'objectif est de connaître ce qu'il y a dans la boite, donc de savoir quelle hypothèse est la bonne parmi ces six-là (...) et bien il faudrait quoi comme précision pour pouvoir trouver la bonne hypothèse ?"*

E9 : *"Six ? Je ne sais pas je ne comprends pas le terme (précision), si je sais ce que ça veut dire mais comment est-ce qu'on peut définir à quel point on veut être précis ?"*

E8 :*"Déjà c'est parmi six possibilités c'est vrai, et pas cinq, on a l'impression qu'il y a cinq... Maintenant j'ai envie de dire qu'il faudrait savoir le nombre de personnes sur lesquelles on fait... Même si je sais que la boîte c'est le nombre de personnes...*

E8 :*"En fait moi j'essaie juste de retrouver la formule que j'avais vue en (rhéto) parce que je suis certaine d'avoir déjà fait cet exercice ou un truc dans le genre... "*

E9 : *"Non, moi je veux quand même (comprendre) ce truc de précision, il nous manque une info, à quel point est-ce qu'on veut être précis. Et ça je ne comprends pas comment est-ce qu'on peut nous même déterminer à quel point on veut être précis. C'est pas quand on met l'expérience qu'on dit 'je veux une précision de autant' ?"*

I :*"Si tu as raison, mais ici les caractéristiques de l'expérience font qu'on peut peut-être déduire la précision dont on aurait besoin".*

E9 : *"Bin de 20 en 20 % ? (voilà, on n'a pas besoin d'être précis à 1 % près)."*

E9 : *"(rappel de situations proches en rhéto), d'un côté je me dis qu'on arrivait à toutes ces formules par un raisonnement (clairement), donc autant qu'on essaie de raisonner"*

E8 :*"Bin regarde un coup dans les exemples ici. (mais il y en a tellement) mais non il y en a 32."*

E9 : *"Mais attends, en gros t'as une population où au plus tu fais des tirages au plus tu seras précis, seulement on veut une précision de 20 % (...). Imaginons on a 100 personnes et on veut une précision de 20 % (et on veut savoir la proportion). (...) Oui on veut la connaitre après mais ce qu'on nous demande c'est pas la proportion du truc. Ce qu'on nous demande c'est combien de prélèvements il faut qu'on fasse pour connaitre ... avec une précision de 20 %."*

E9 : *"Mais je pense que directement avec notre petite formule, je ne sais pas si on y arrivera."*

E9 : *"(partons d'une population de $n = 100$ et dans laquelle on fait des tirages, sachant qu'il faut une précision de 20 %). Mais du*

coup j'arrive à un stade où il faudrait connaître le nombre de personnes dans cette population pour savoir combien de tirages je vais faire parce qu'au plus elle est grande au plus je vais devoir faire de tirages pour être précis à 20 %, ou pas ?"

6. **Réalisation de tirages**

49', l'investigateur suggère de partir sur des tirages expérimentaux.

I : *"Si vous voulez vous pouvez faire des tirages (ok essayons)."*

E : $X = 5$, $N = 10$

E8 : *"On a une chance sur deux de tomber (. . .) sur la blanche ou sur la croix"*

E9 : *"Non vu qu'il y en a cinq dedans, c'est pas un pile ou face."*

E8 : *"(demande aide de l'investigateur) Parce qu'en fait on se rend compte qu'on a une chance sur cinq de tomber. . . il y a cinq boules dans le. . . mais chaque boule elle a une chance sur deux d'être soit autre ou soit H3N2. (qu'est-ce qui vous permet de savoir que c'est une chance sur deux ?). Parce qu'on a deux sous-ensembles, on a soit H3N2 soit autres (l'investigateur dément).*

E8 : *"En fait je ne sais pas comment combiner les deux, le fait qu'on ait une chance sur cinq de tomber sur cette boule-là et que cette boule-là a deux possibilités d'être soit autres ou soit H3N2."*

E8 : *"Et par exemple ici on a fait 10 prélèvements et on est tombés sur une chance sur deux (50 % de H3N2)."*

E9 : *"Or ça ($p = 0,5$) c'est pas ce qu'on est censé obtenir".*

I : *(il y a entre 1 et 4 H3N2 dans la boite puisque les deux sous-types ont été observés)*

I : *"Et si on refaisait 10 prélèvements, vous pensez qu'on serait loin de. . . ($P = 0,5$) ?"*

E9 : *"Bah, peut-être, on n'en sait rien, bon on réessaie ?"*

E8 : *"Mais c'est tellement aléatoire, je ne sais pas on peut tomber cinq fois sur la même boule et"*

E9 : *"Comme dix fois, imaginons on tomberait 10 fois sur H3N2, si on n'a pas fait ces dix autres (les 10 premiers, $P = 0,5$) prélèvements on peut croire qu'il n'y a pas d'autres. . . petites boules (sous-type)."*

E9 : *"Avec 10 prélèvements on a une précision de. . . pas beaucoup, mais sachant qu'on a que 5 petites boules."*

E8 : *"Moi ça m'ennuie les cinq boules, ça me perturbe."*

I : *(analogie entre boite et pièce de monnaie)*

E8 : *(cherche à appliquer une formule, cherche la correspondance entre les termes vus dans les formules et les éléments de l'énoncé). "Moi j'ai envie d'une formule, j'ai envie d'une formule (rires)."*

E9 : *"Moi j'ai l'impression qu'il n'y a pas de formules pondues pour ça (Si). Bin surement mais pas à notre connaissance (mais peut-être pas en une formule juste une)."*

7. **Suggestion de raisonnement hypothético-déductif**

1h02, l'investigateur tente d'amener les étudiantes à considérer les résultats possibles sous chacune des hypothèses.

I : *"Et si là-dedans il y avait 1 H3N2, est-ce que c'était possible de tomber sur 5/10 ?"*

E9 : *"Bah, possible oui"*

E8 : *"En soi c'est dû au hasard quoi, parce qu'on peut tomber 10 fois sur la même boule tout comme on peut ne jamais tomber dessus."*

I : *"Et on ne sait pas mesurer à quel point, s'il n'y avait que une seule H3N2, si le résultat qu'on a obtenu-là il était possible ? Il était facile à obtenir ?"*

E9 : *"Si il n'y en avait qu'une, (il ne serait) pas facile à obtenir mais possible ça oui, tout est possible). Mais je ne vois pas la suite de ce raisonnement."*

I : *"On ne sait pas le mesurer ? (comment ça ?) A quel point c'était difficile d'obtenir cinq sur 10 (si $\pi = 20$ %) ?"*

E9 : *"Bin ici comme on a eu 50-50, on peut penser qu'il y a 3 et 2 ou 2 et 3 ($\pi = 40$ % ou $\pi = 60$ %)."*

I : *"Mais est-ce qu'avec 4 et 1 c'était possible aussi ?"*

E9 : *"Bin si, c'est là le problème (rires). C'est possible, c'est moins possible mais c'est possible."*

E9 : *"Mais après je ne comprends pas comment on peut arriver à calculer juste avec des suppositions.*

E8 : *"(Cherche toujours une formule). Est-ce que dans le livre ce genre d'exemple est expliqué ?".*

E : *(Toujours en train de chercher avec des probas)*

I : *"En fait il faut oublier qu'il y a cinq boules dedans. Le fait qu'il y ait cinq boules dedans nous sert juste à dire que les proportions possibles, qu'on va trouver, sont limitées, c'est juste 0, 1/5, ... Mais pour le reste, ce n'est jamais qu'un truc qui va donner des résultats avec une certaine probabilité. Comme si on avait une pièce et qu'on disait 'est-ce que ma pièce elle est truquée ou pas ?', il faut lancer combien de fois, lancer combien de pile ou face pour savoir si ma pièce est équilibrée ou truquée ? (Mais ça c'est ce qu'on nous demande au final). Oui c'est la même chose, j'aurais pu demander ça aussi."*

E9 : *"Du coup, ce qu'on cherche à savoir c'est : combien, là dans la petite boite, il y a de boules H3N2. (oui) Mais ça, entre guillemets, on ne pourra jamais le savoir vraiment. On ne peut que s'en rapprocher fortement... avec nos probabilités (qu'est-ce que tu appelles nos probabilités ?). Bin pour moi c'est ce fameux 1/10 (je crois que que le 1/10 n'est pas). Mais le 1/10 n'est pas correct mais, vu qu'au final... enfin... "*

I : *"Et ici, si on recommençait les 10 tirages ?"*

E9 : *"Mais imaginons (qu'on recommence les 10 tirages), au final on peut tomber sur n'importe quoi... Que ce soit ... parce qu'ici on est tombé sur 1/2 - 1/2."*

E8 : *"Est-ce qu'on devrait essayer de trouver les possibilités, genre est-ce qu'on devrait se dire (qu') on a la possibilité de les avoir toutes blanches, 1-4, ... Et essayer de faire ça et puis... comme on fait avec les possibilités d'une pièce avec pile ou face. (ça peut, c'est une possibilité)."*

E9 : *"Ha oui, ok, mais seulement avec tout ça je ne sais pas où ça nous mène."*

8. **Approche par les simulations**

01h11, l'investigateur propose d'utiliser des simulations via un tableur pour arriver à la distribution des résultats attendue sous une hypothèse.

I :*"(. . .) il faut juste qu'on lui (logiciel) dise quelle est la probabilité".*

E9 : *"Mais ça. . . donc on doit la trouver déjà à l'avance la probabilité".*

I : *"ou alors on la pose. On dit : si c'était 1/5".*

E9 : *"Ah, on a le droit de la poser ? ! Ou alors on fait pour si cétait 1/5, si c'était 2/5, si c'était 3/5, . . . et puis on vérifie."*

I :*"Si c'était 1/5, et si on en prenait 10, on aurait quoi comme résultats ?"*

I :*"On peut voir quel est l'ensemble des résultats qu'on pourrait obtenir dans ces conditions-là (mais ici on les obtient tous en fait). Non par exemple 7, 8, 9, 10, on ne les a jamais vus."*

E8 :*"Donc sur 10 fois qu'on fait, en posant la condition qu'il y a une boule sur cinq qui contient le H3N2, on arrive parfois à tirer 6. Mais je ne sais pas comment expliquer ça."*

E9 : *"Là on a posé la condition qu'il y a une boule sur cinq ? (oui) Haa ok."*

I :*"On a dit : si on faisait 10 tirages et si on avait une chance sur cinq d'avoir un H3N2, on pourrait avoir quoi comme résultats ?"*

E9 : *"Le 0, 1, 2, 3 ce sont ceux qui reviennent le plus souvent."*

E8 :*"Mais avec ça qu'est-ce qu'on peut en conclure ? Parce qu'après on va quand même tester la même chose avec le 2/5, puis le 3/5 puis 4/5. . . Puis on va comparer les données et puis c'est à quel moment qu'on se dira 'oui ça c'est le bon pourcentage !' ?"*

I :*"Là les résultats possibles ils varient dans quelle fourchette ?"*

E9 : *" 0-6 (oui, 0-6 là tu les as tous, mais si tu veux les avoir quasi tous ?) 0-3 ? (voilà, 0-3, 0-4)."*

I :*"0 et 3 sur 10, ça fait quelle proportion ?"*

I :*"En fait, ici, j'essaie de vous faire mesurer quelle est la précision qu'on peut avoir (ha, ça c'est la précision ?). Il y a moyen de voir, avec des simulations comme ça, si ,quand le résultat est 1/5, si on est souvent proche de 1/5, ou combien de fois on va. . . à quel point on s'en écarte."*

E8 :*"Oui, ça fluctue, mais on va utiliser comment ces données-là par rapport à. . . "*

E8 :*"Ici c'est le 2-3-4-5 qui l'emporte, enfin qui ont les plus gros. . . mais je ne sais pas comment utiliser ces infos. . . "*

I :*"on va essayer autre chose, on va abandonner cette idée-là".*

9. **Intervalles de décision déduits de la distribution binomiale**

1h27, l'investigateur fait le lien entre les résultats simulés par Excel et la distribution binomiale et réexplique la lecture tables binomiales.

E8 :*"Ok, là on a des sortes de petites plages. Et avec ça comment est-ce qu'on va déterminer qu'en fait on est à 1/5, 2/5, 3/5, . . . ?"*

I :*" Et bien, si (. . .) maintenant on fait des tirages et qu'on obtient 0. Et bien le résultat il est compatible avec l'hypothèse qui dit qu'il y a 0 H3N2, mais aussi avec celle-là (1/5) mais pas avec celle-là (2/5) et pas avec les autres. Si on obtient 1, ce n'est plus compatible avec l'hypothèse qui dit qu'il n'y en n'a pas (H_0), c'est compatible avec celle-ci (H_1), . . . "*

E9 : *"Ca c'est ce qu'on a dit nous, mais au final pourquoi est-ce que ce ne serait pas compatible ? C'est juste que nous on a défini les limites comme ça."*

E8 :*"Maintenant, quand vous dites on relance, ça veut dire qu'on ferait nous-mêmes comme ça (un tirage) ?"*

I :*"Oui, et à partir du résultat on pourrait voir quelles sont les compositions compatibles"*

I :*"Par exemple ici on en a fait 10 (tirages) et on est tombés sur 5 (H3N2). Donc par exemple 5, c'est pas vraiment compatible avec cette hypothèse-ci (H_1) par contre c'est compatible avec 0,4 (H_2) et surement avec 0,6 parce que c'est au milieu, mais peut-être pas avec 0,8 parce que si c'est symétrique, si 0,2 n'est pas compatible, j'imagine que 0,8 non plus."*

E8 :*"Ok, là on part sur... on a déterminé que c'est soit 2/5 soit 3/5 des boules qui possèdent (H2N3)."*

I :*"Mais l'ennui c'est qu'on ne sait pas définir clairement, on ne sait pas trancher ici. (Et c'est là que les 20 % près vont nous dire... c'est ça qui va trancher ?). A peu près oui, disons que, tant qu'il y a des résultats qui nous permettent pas de trancher, c'est qu'on a peut-être pas assez de tirages."*

I :*"Et si on faisait 20 tirages, est-ce qu'il y aurait des résultats qui seraient ambigus ?"*

E8 :*(Visiblement a compris quels résultats sont ambigus)*

E8 :*En fait, plus on va en faire, plus ça va être facile de déterminer (π ?)... plus on va avoir des grands écarts...*

E8 :*(Discussion sur la règle à appliquer pour considérer les résultats probables sous une hypothèse)*

E8 :*"Ici je n'ai même plus d'ambiguité, (...) Moi j'ai envie de dire qu'il faut même aller jusqu'à faire 200 prélèvements pour être super précis."*

E9 : *"Non parce qu'on veut une précision de 20 %, et on ne veut pas forcément plus ni forcément moins (ha c'est ça)."*

E8 :*"Et la ici ma précision de 20 %, comment est-ce que je peux regarder ?"*

I :*"Bin si tu arrives à séparer sans ambiguité deux hypothèses qui sont éloignées de 20 %... "*

E9 : *"C'est qu'il faut ce nombre de tirages-là (donc 50 c'est bon, ici dans ce cas)."*

E9 : *"Bin, chez nous (dans le cas de la boite) du coup ce serait bon aussi ?"*

E8 :*"Si on fait à 100, là on va avoir un écart entre les deux de plus que 20 % (trop grande précision avec $N = 100$). Là j'ai l'impression qu'on est tout pile dans la précision. (Effectivement au plus on va augmenter (N) au plus ces résultats-là vont se séparer)."*

E8 :*"Maintenant, je ne vais pas le cacher, c'est pas du tout comme ça que je pensais qu'on allait ... (résoudre le problème)".*

I :*"Oui, je sais que je vous ai un peu orienté sur la fin."*

E8 :*"Et mais même, jamais je n'aurais eu l'idée d'aller faire ça ! Moi, pour moi, une petite formule ça allait être (suffisant). taper mon chiffre, mes possibilités et (appliquer la formule et avoir la réponse)."*

E9 : *"Mais c'est ça notre réponse c'est quand même bien 50 au final ?"*

10. **Lien entre taille d'échantillon et probabilité d'erreur**

1h40, l'investigateur amène à considérer le risque d'erreur.

I :*"Oui, mais il manque encore un élément de nuance (...)".*

E8 :*"On n'a pas trouvé en fait notre fraction pour le moment ? On ne sait pas si c'est 0,2, 0,4 ou 0,6... "*

I :*"Et bien on peut essayer. Si on a dit que 50 (tirages) sont suffisants pour savoir... (maintenant on a le chiffre) On en a déjà fait 10, il n'y a plus qu'à en faire 40 autres."*

E8 : *"Oui, mais encore une fois, là on va faire 50 fois, mais c'est une session de 50 fois, si on en fait encore d'autres, les réponses seront toujours différentes. On ne sera jamais sûres en fait."*

I : *"Non c'est vrai, tu ne seras jamais sûre. Tu pourras toujours dire, en fait il y avait... très peu dans la population ont le H3N2 mais tous ceux que j'ai tiré, par malchance, ils l'avaient. Et ça tu n'en seras jamais sûre. Il faudrait les examiner tous pour savoir."*

E8 : *"On a cinq médecins dans le monde entier, ils vont faire la même expérience, ils vont tous tomber dans possibilités différentes et ils vont tous poser un diagnostic différent... Enfin c'est un problème, c'est pas universel, tout le monde aura un truc (résultat observé) différent."*

I : *"Oui, sauf que, au plus ils vont faire de mesures au plus cela va se rapprocher... "*

I : *"Mais il manque un petit élément (dans la démarche jusqu'ici) c'est que, en faisant ça on peut se tromper. Parce qu'on dit :quand c'est cette hypothèse-là qui est vraie, on va observer ces résultats-là. Mais ce n'est pas tout à fait tout à fait vrai. (Oui, on peut toujours tomber là-dessus ou là-dessus...). Et il y a quelle chance que ça tombe là-dessus ou là-dessus ? (résultats improbables sous H). Ou quelle chance que ça tombe là-dedans (zone d'acceptation). Vu qu'il y a 94 % de chances d'avoir entre 0 et 14... (Bin on a 6 % de chances de tomber là-dessus ou là-dessus). On a 6 % de chances de tomber au dessus, et là en-dessous ? (là on a 0,4, heu 4 %). Voilà. Du coup on a 10 % de chances de tomber en dehors et 90 % de chances de tomber dedans."*

E9 : *"Du coup ça nous fait perdre un peu de précision quand même"*

I : *(Il faut distinguer le risque d'erreur et la précision)*

E9 : *"Donc au final, quelque part c'est ... enfin c'est pas toutes les réponses possibles mais en dépendant de... (c'est subjectif) oui c'est ça (c'est nous qui avons choisi ça donc). Si moi j'ai envie de prendre des valeurs qui sont plutôt comme ça ou comme ça, bin du coup j'irai plus loin, donc c'est en fonction de comment tu justifies la suite quoi plutôt."*

I : *"En fait toutes les réponses étaient bonnes, c'est juste que chacune est liée à un risque d'erreur différent."*

E9 : *"Mais ça veut dire que c'est subjectif, peut-être que nous on va trouver quelque chose, qu'un autre groupe aura trouvé ... (autre chose)."*

I : *"Si on accepte tel risque d'erreur, alors le bon nombre ce serait autant."*

E8 : *"Donc en fait la réponse ce serait pas forcément 50, c'est nous qui avons 50 mais on le justifie parce qu'on a décidé de prendre ça et qu'on a X % d'erreur (qu'on a décidé de choisir 10 % d'erreur)."*

E9 : *"Mais du coup, maintenant, qu'on dit ça, nos petits 20 % de précision, ils intervenaient où encore ?"*

I : *(explique que pour la même taille d'échantillon, une précision de 10 % n'aurait pas été obtenue, car des résultats auraient été ambigus)*

E8 : *"Et maintenant, je ne sais pas si on a le temps mais on ne va pas refaire cinquante fois l'expérience, mais comment est-ce qu'on va savoir pour cet exercice que c'est 0,4 ou 0,6 ?"*

I : *(une fois que la méthode est au point, que l'on a fixé le risque d'erreur et déterminé le N, il ne reste qu'à faire les tirages et tirer la conclusion).*

E9 : *"Une distribution binomiale en fait c'est exactement ce qu'on a fait sur Excel mais déjà fait à l'avance (oui et avec une infinité de tirages)."*

11. **Résumé**

1h58, une étudiante résume la démarche.

E8 :*"Donc, on cherche le nombre de prélèvements / nombre de tirages qu'il faut pour analyser la proportion d'H3N2 dans une population. La méthode qu'on a choisie finalement c'est par Excel / en regardant une table déjà prédéfinie des distributions binomiales. (quand tu dis slash, c'est 'ou'?) . Oui, donc (les tables binomiales sont faites sur une infinité de tirages). C'est un peu des chiffres déjà pré... On tomberait toujours sur ces chiffres si on prenait une infinité de tirages. Et c'est un pourcentage ... Le pourcentage est déjà calculé en fait. C'est comme si on avait fait par Excel mais c'est déjà fait pour nous. On a regardé le tableau de (N =) 10, parce qu'on avait déjà auparavant fait des tirages de 10, on était tombé sur 5/10 de H3N2. Donc d'abord on regarde les proportions de 20 % de précision, vu qu'on a que cinq boules, donc ... et qu'on élimine le 0 % et le 100 % parce que quand nous on a fait l'expérience on était tombé sur au moins une boule blanche et au moins une boule avec des croix. Donc on regarde seulement les colonnes de 0,2, 0,4, 0,6 et 0,8. On a ensuite choisi une plage entre 10 et 90 (%) en acceptant une marge de 10 % de risque avant et après. Et ensuite on s'est rendu compte qu'il y avait des ambiguïtés à des endroits où ça pouvait être plusieurs possibilités (hypothèses) encore. On a refait avec ($N = 20$). Avec 20 il y avait encore des ambiguïtés et quand on est passée à 50, là on avait des plages qui fluctuent entre (le quantile) 10 et 90 (%) bien définies, donc là il n'y avait plus d'ambiguïté. Et donc là quand on recommence nos tirages expérimentaux nous-mêmes, en acceptant 10 % d'erreur, on se dit que le nombre qu'on aura tombera sur le pourcentage qu'il y a de boules à l'intérieur."*

E8 :*"Même moi (qui) explique je me dis que c'est un peu du chinois quoi!"*

I :*"Ok merci"*

E8 :*"En même temps je pense qu'on n'aurait pas travaillé comme ça si on n'était pas juste à deux avec l'enregistrement." (Elles n'auraient pas persévéré).*

3.4.5 Séance 3 : analyse sémantique

La description chronologique du déroulement des séances enregistrées fait apparaitre que les raisonnements des étudiants ne correspondent pas aux attentes de l'investigateur. On peut penser que cela devrait se refléter dans les concepts manipulés par les uns et les autres.

Afin d'objectiver les écarts entre les concepts mobilisés par les uns et les autres, nous proposons une analyse sémantique des enregistrements. L'idée est d'identifier les concepts que les étudiants utilisent bien plus souvent, ou au contraire plus rarement, que l'investigateur.

Méthode

Pour cela nous avons suivi la méthodologie suivante.

Les verbatim des échanges entre étudiants ou entre les étudiants et l'investigateur représentent 584 interventions pour 71893 mots (dont 26875 pour l'investigateur et 45018 pour les étudiants). L'identification des concepts dans ces verbatim a nécessité les étapes suivantes :

1. **Nettoyage des caractères** : retrait des espaces, de la ponctuation, remplacement des caractères majuscules par des minuscules et remplacement des caractères accentués ou spéciaux.
2. **Suppression des termes courants** (dont les pronoms personnels, les déterminants, conjugaisons du verbe avoir et être). Pour cela nous avons utilisé la liste de *stopwords* (que l'on peut traduire par *mots vides*) fournie avec le package *tm* (pour *Text Mining*) disponible pour le logiciel R 3.6.0.
3. **Remplacement des mots par leur racine**. Les mots *probable* et *probables*, par exemple, sont remplacés par la racine *probabl*. Cela se fait à l'aide de l'algorithme de Porter mobilisé dans la fonction *stemDocument* du package *tm*. On obtient 821 racines.
4. **Association des racines à un sens**. Cette opération nécessite de regarder, pour chaque racine, les extraits dans lesquels elle apparait afin d'en déterminer manuellement le sens auquel l'associer. La racine *certain*, par exemple, aurait pu être associée au concept de *certitude* mais l'examen du contexte dans lequel elle apparait indique qu'elle est essentiellement utilisée pour désigner "certains objets" plutôt que "des choses certaines". La racine *certain* n'a donc pas été associée à un concept particulier. La racine *chinois*, par exemple, intervient dans l'expression "c'est du chinois" et a été associée au concept de *difficulté*.

   Ce travail conduit à l'identification de 61 concepts qui seront séparés en deux groupes : les verbes ($n = 27$) et les autres concepts ($n = 34$) [17].
5. **Suppression des concepts trop rares**. Les 10 concepts mobilisés moins de 5 fois dans l'ensemble des verbatim sont écartés de l'analyse. Il reste donc 51 concepts dont 17 verbes et 34 autres concepts.

Nous avons ensuite calculé les fréquences relatives d'utilisation de chaque concept séparément pour les étudiants et pour l'investigateur. Prenons le verbe *devoir*, par exemple (voir tableau 3.11). Il est mobilisé 3 fois par l'investigateur en 26875 mots, ce qui donne $1,1$ utilisations tous les 10000 mots ($FR_{investigateur}$). De leur côté, les étudiants mobilisent 35 fois le concept en 45018 mots soit $7,8$ utilisations tous les 10000 mots ($FR_{etudiants}$). Le rapport $\frac{FR_{etudiants}}{FR_{investigateur}}$ est de $7,0$ signifiant que les étudiants utilisent 7 fois plus le verbe *devoir* que

17. La distinction entre verbes et concepts n'est pas toujours évidente. Certains choix peuvent être discutables mais cela ne modifie en rien l'analyse qui est faite car l'intérêt de cette distinction est simplement de diviser une importante masse de concepts en deux séries intelligibles.

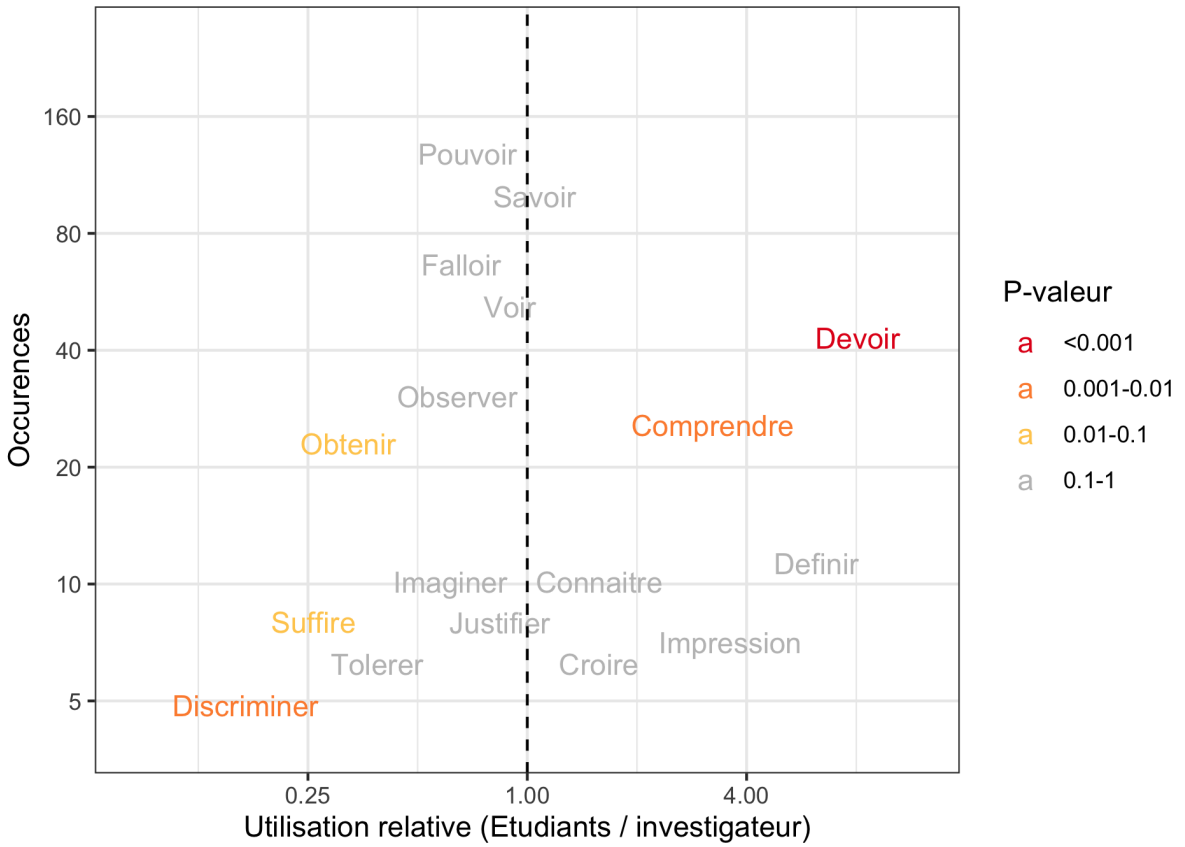


FIGURE 3.14 – **Représentation de l'utilisation relative des verbes entre les étudiants et l'investigateur**. L'axe des abscisses représente l'utilisation relative du concept définie comme étant $\frac{FR_{etudiants}}{FR_{investigateur}}$. L'axe des ordonnées représente le nombre de fois que le concept est mobilisé par les étudiants et l'investigateur. Pour les deux axes, un échelle logarithmique est utilisée. Le code couleur donne la significativité statistique de ces résultats. Les utilisations relatives supérieures à 10 ou inférieures à 1/10 ont été ramenées respectivement à 10 et 1/10 afin de permettre leur visualisation sur le graphique.

l'investigateur (Intervalle de confiance à 95 % sur ce ratio : [2,2 à 35,4], $P < 0.0001$ [18] [19]).

Résultats

Le résultat de cette analyse sont résumés dans les tableaux 3.11 et 3.12 et les figures 3.14 et 3.15.

On constate que des verbes tels que *pouvoir*, *falloir*, *savoir* ou *voir* sont autant utilisés

18. Une telle différence d'utilisation se trouve être très incompatible avec l'hypothèse stipulant qu'au niveau de la population, étudiants et investigateur utilisent autant le concept d'*hypothèse*. Dans le cadre d'une analyse exploratoire comme celle-ci, en dehors de toute hypothèse préalable, difficile de tester une autre hypothèse que l'hypothèse nulle.

19. Intervalle de confiance et P-valeur obtenus par un test binomial.

TABLE 3.11 – **Comparaison de l'utilisation des principaux verbes par les étudiants et l'investigateur**. FqInv et FqEt : nombre de fois où le concept est mobilisé par l'investigateur et par les étudiants, FrInv et FrEt : fréquence relative du concept dans le discours de l'investigateur et des étudiants (par 10000 mots), RR : FrEt/FrInv, low et up : limites inférieure et supérieure de l'intervalle de confiance à 95 % sur le RR, Pval : P-valeur.

	Verbe	FqInv	FqEt	FrInv	FrEt	RR	low	up	Pval
1	Devoir	3	35	1.1	7.8	6.965	2.196	35.395	6.5e-05
2	Comprendre	4	25	1.5	5.6	3.731	1.288	14.752	7.0e-03
3	Discriminer	5	0	1.9	0.0	0.000	0.000	0.651	7.3e-03
4	Obtenir	16	10	6.0	2.2	0.373	0.151	0.874	1.4e-02
5	Suffire	6	3	2.2	0.7	0.298	0.048	1.398	8.8e-02
6	Definir	1	9	0.4	2.0	5.373	0.745	235.497	1.0e-01
7	Observer	14	13	5.2	2.9	0.554	0.240	1.272	1.6e-01
8	Falloir	33	42	12.3	9.3	0.760	0.470	1.237	2.4e-01
9	Pouvoir	49	65	18.2	14.4	0.792	0.538	1.172	2.5e-01
10	Impression	1	7	0.4	1.6	4.179	0.537	188.338	2.7e-01
11	Tolerer	4	3	1.5	0.7	0.448	0.066	2.647	4.4e-01
12	Voir	20	26	7.4	5.8	0.776	0.417	1.466	4.5e-01
13	Savoir	35	53	13.0	11.8	0.904	0.579	1.428	6.6e-01
14	Imaginer	4	5	1.5	1.1	0.746	0.161	3.761	7.4e-01
15	Justifier	4	5	1.5	1.1	0.746	0.161	3.761	7.4e-01
16	Connaitre	3	6	1.1	1.3	1.194	0.255	7.378	1.0e+00
17	Croire	2	5	0.7	1.1	1.492	0.244	15.673	1.0e+00

par les étudiants que par l'investigateur. Par contre, l'investigateur utilise bien plus le verbe *discriminer* ($RR = 0,00$, [0,00 à 0,65], $P = 0,007$) tandis que les étudiants emploient $3,8$ fois plus le verbe *comprendre* ([1,3 à 14,8], $P = 0,007$) et $7,0$ fois plus le verbe *devoir* ([2,2 à 35,4], $P < 0.0001$).

Concernant les autres concepts, l'analyse sémantique révèle que les éléments liés au *contexte* de la situation (virus, malade, médecin, patients, *etc.*) ou au *dispositif* (boîte, balles, *etc.*) sont assez équitablement répartis entre l'investigateur et les étudiants. C'est également le cas de concepts tels que la *proportion*, la *probabilité*, l'*échantillon*, *etc.* Par contre l'investigateur utilise bien plus souvent les concepts d'*hypothèse* ($RR = 0,053$ [0,010 à 0,17], $P < 0,0001$), de *résultat* ($RR = 0,16$ [0,078 à 0,31], $P < 0,0001$), de *quantification* ($RR = 0,07$ [0,001 à 0,48], $P = 0,001$) ou encore de *compatibilité* ($RR = 0,05$ [0,001 à 0,34], $P = 0,00006$). Les étudiants, quant à eux, utilisent plus fréquemment les concepts de *formule* ($RR = 4,5$ [1,6 à 17,5], $P = 0,001$), les termes techniques propres aux distributions théoriques discrètes (table, succès, échec, distribution, indépendance) telles que la distribution *binomiale* ($RR = 2,2$ [1,1

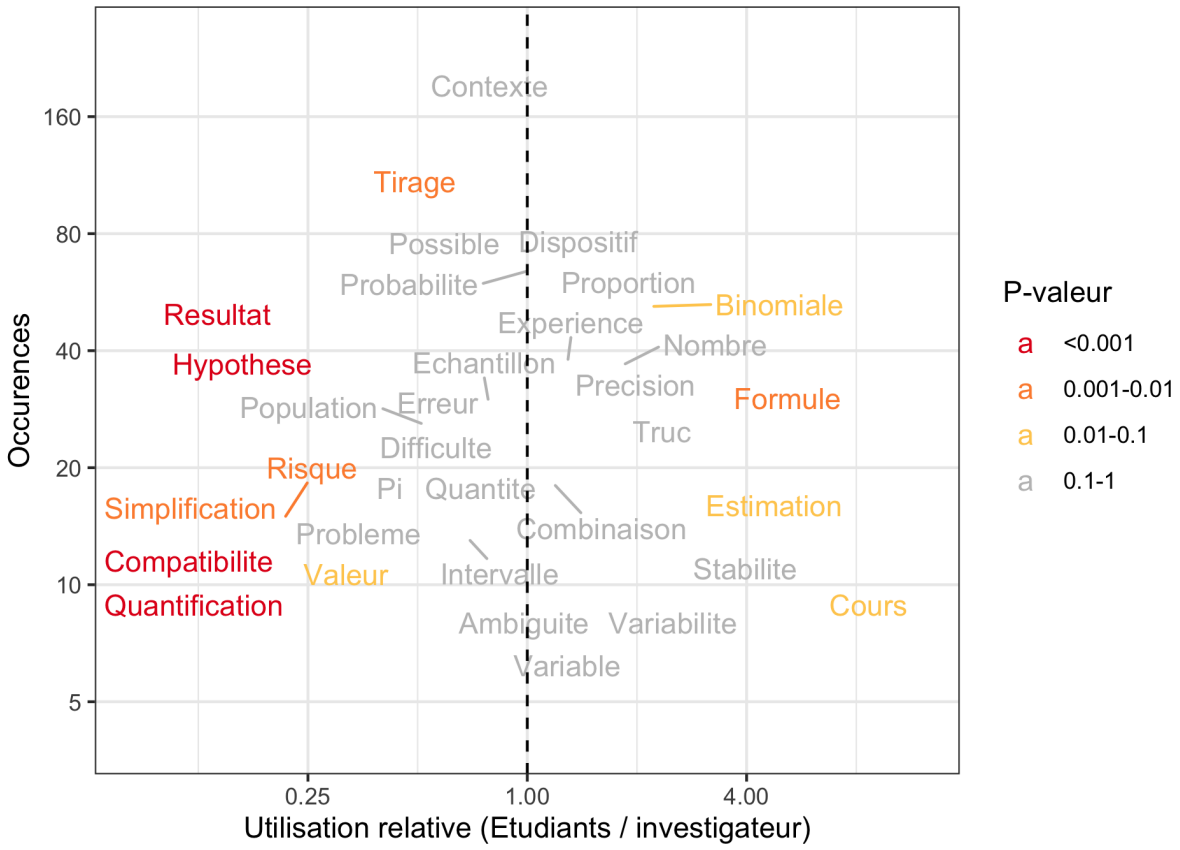


FIGURE 3.15 – **Représentation de l'utilisation relative des concepts entre les étudiants et l'investigateur**. L'axe des abscisses représente l'utilisation relative du concept définie comme étant $\frac{FR_{etudiants}}{FR_{investigateur}}$. L'axe des ordonnées représente le nombre de fois que le concept est mobilisé par les étudiants et l'investigateur. Pour les deux axes, un échelle logarithmique est utilisée. Le code couleur donne la significativité statistique de ces résultats.Les utilisations relatives supérieures à 10 ou inférieures à 1/10 ont été ramenées respectivement à 10 et 1/10 afin de permettre leur visualisation sur le graphique.

à 4,8], $P = 0,02$) et font plus souvent référence au *cours* théorique ($RR = +\infty$ [1,3 à $+\infty$], $P = 0.02$).

TABLE 3.12 – **Comparaison de l'utilisation des principaux concepts par les étudiants et l'investigateur**. FqInv et FqEt : nombre de fois où le concept est mobilisé par l'investigateur et par les étudiants, FrInv et FrEt : fréquence relative du concept dans le discours de l'investigateur et des étudiants (par 10000 mots), RR : FrEt/FrInv, low et up : limites inférieure et supérieure de l'intervalle de confiance à 95 % sur le RR, Pval : *P*-valeur.

	Concept	FqInv	FqEt	FrInv	FrEt	RR	low	up	Pval
1	Hypothese	34	3	12.7	0.7	0.053	0.010	0.167	5.9e-12
2	Resultat	44	12	16.4	2.7	0.163	0.078	0.314	5.2e-10
3	Compatibilite	12	1	4.5	0.2	0.050	0.001	0.336	6.3e-05
4	Quantification	9	1	3.3	0.2	0.066	0.002	0.479	9.5e-04
5	Formule	4	30	1.5	6.7	4.477	1.578	17.493	1.2e-03
6	Tirage	63	60	23.4	13.3	0.569	0.392	0.823	2.0e-03
7	Simplification	11	3	4.1	0.7	0.163	0.029	0.616	3.5e-03
8	Risque	11	4	4.1	0.9	0.217	0.050	0.733	6.0e-03
9	Binomiale	11	41	4.1	9.1	2.225	1.123	4.800	1.5e-02
10	Cours	0	10	0.0	2.2	Inf	1.338	Inf	1.7e-02
11	Valeur	8	4	3.0	0.9	0.298	0.066	1.114	6.8e-02
12	Estimation	3	15	1.1	3.3	2.985	0.844	16.085	8.8e-02
13	Population	14	12	5.2	2.7	0.512	0.216	1.192	1.0e-01
14	Pi	11	9	4.1	2.0	0.488	0.179	1.297	1.1e-01
15	Nombre	9	28	3.3	6.2	1.857	0.852	4.474	1.3e-01
16	Truc	5	17	1.9	3.8	2.030	0.719	7.036	1.9e-01
17	Possible	30	37	11.2	8.2	0.736	0.443	1.234	2.1e-01
18	Stabilite	2	9	0.7	2.0	2.686	0.556	25.551	2.3e-01
19	Probleme	8	7	3.0	1.6	0.522	0.161	1.648	2.8e-01
20	Variabilite	2	7	0.7	1.6	2.089	0.398	20.614	5.0e-01
21	Experience	12	26	4.5	5.8	1.293	0.630	2.814	5.1e-01
22	Contexte	86	131	32.0	29.1	0.909	0.688	1.208	5.3e-01
23	Erreur	11	14	4.1	3.1	0.760	0.320	1.849	5.4e-01
24	Precision	9	20	3.3	4.4	1.327	0.577	3.309	5.7e-01
25	Intervalle	6	7	2.2	1.6	0.696	0.200	2.509	5.7e-01
26	Echantillon	13	17	4.8	3.8	0.781	0.357	1.748	5.7e-01
27	Dispositif	22	43	8.2	9.6	1.167	0.683	2.048	6.1e-01
28	Proportion	22	43	8.2	9.6	1.167	0.683	2.048	6.1e-01
29	Combinaison	6	12	2.2	2.7	1.194	0.415	3.877	8.1e-01
30	Difficulte	9	14	3.3	3.1	0.929	0.374	2.432	8.3e-01
31	Ambiguite	3	6	1.1	1.3	1.194	0.255	7.378	1.0e+00
32	Probabilite	24	40	8.9	8.9	0.995	0.585	1.725	1.0e+00
33	Quantite	7	13	2.6	2.9	1.109	0.411	3.282	1.0e+00
34	Variable	2	5	0.7	1.1	1.492	0.244	15.673	1.0e+00

3.4.6 Séance 4

Une partie de la quatrième séance de travaux pratique a finalement été consacrée à la finalisation du travail entamé lors de la séance précédente sur la situation fondamentale du test d'hypothèses.

Ensuite, certains groupes d'étudiants ont présenté leur raisonnement devant les autres groupes. L'assistant a ensuite pris le relais pour faire le lien entre les raisonnements ainsi présentés et l'une ou l'autre des pistes de résolution possibles.

Dans l'ensemble, on peut remarquer que les étudiants acceptent la procédure présentée par l'assistant, peuvent en comprendre chacune des étapes mais celle-ci leur semble assez peu intuitive.

Par ailleurs, faire le lien entre les connaissances qui ont ainsi émergé et les notions formelles d'hypothèse nulle et alternative, de zones d'acceptation , d'erreurs α et β ne s'est pas toujours révélé facile.

3.5 Analyse *a posteriori*

> "*[Dans la méthodologie de l'ingénierie didactique, la phase d'expérimentation] est suivie d'une phase d'analyse dite a posteriori qui s'appuie sur l'ensemble des données recueillies lors de l'expérimentation : observations réalisées des séances d'enseignement mais aussi productions des élèves en classe ou hors classe. Ces données sont souvent complétées par des données obtenues par l'utilisation de méthodologies externes : questionnaires, entretiens individuels ou en petits groupes*". [Artigue, 1988]

Après avoir décrit les observations de manière chronologique et relativement détaillée, nous allons tenter d'analyser le déroulement des quatre séances de travaux pratiques avec la grille de lecture que nous fournit la théorie des situations didactiques.

3.5.1 Séance 1

Lors de la première séance de travaux pratiques, les observations suggèrent que les étudiants ont globalement accepté la dévolution proposée par l'enseignant.

Ils ont globalement bien compris les consignes (comparer les trois séries de données à l'aide d'un tableau résumé puis d'un graphique) et ont, dans l'ensemble, accepté de mettre à l'épreuve leurs conceptions. Face à leurs premières réponses, ils ont pu se rendre compte, seuls ou à l'aide des questionnements de l'assistant, du fait que certaines conceptions initiales (telles que la

moyenne, le minimum ou le maximum) étaient inopérantes dans le cas présent et ont pu sentir le besoin d'en utiliser d'autres.

La plupart des groupes d'étudiants ont construit de nouveaux concepts par eux-mêmes, tandis que certains ont choisi d'appliquer les formules (notamment de variance et d'écart-type) vues dans le formulaire ou dans le syllabus du cours.

Parmi ceux qui ont construit de nouveaux concepts, on peut noter que les trois conceptions suivantes ont fréquemment émergé.

1. Calcul de la **proportion d'individus qui ont perdu du poids** dans les trois séries. Cette mesure apporte effectivement une information supplémentaire et intéressante dans l'idée de comparer trois régimes alimentaires.
2. **Écart-moyen à la moyenne**. Face au besoin de quantifier la dispersion des données, de nombreux groupes d'étudiants se sont lancés dans le calcul d'un écart à la moyenne, soit :

   $$Ecart - moyen = \frac{\sum |x_i - \bar{x}|}{n}$$

   avec x_i : les observations individuelles, $\bar{x}$: la moyenne de la série et n : le nombre d'observations.

   Il s'agit d'une mesure de dispersion adaptée au problème de l'analyse descriptive et qui permet, dans la phase d'institutionnalisation de faire facilement faire le lien avec l'écart-type.

   $$Ecart - type = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

3. Représentation des trois séries de données en **nuage de points** (voir figure 3.16, haut). Il s'agit d'un graphique intéressant dans la mesure ou il permet de retirer des informations assez générales sur les trois distributions à comparer mais relativement imprécis car il ne donne qu'une idée très grossière de la fréquence des différentes classes de valeurs.

   Ce qui explique l'importante prévalence de cette représentation est vraisemblablement l'outil informatique utilisé pour traiter les données : le logiciel Microsoft Excel (version 2010) car il s'agit de la représentation obtenue par défaut lorsque l'on cherche à insérer un graphique en sélectionnant une série de données. De nombreux groupes d'étudiants sont donc passés par-là. Et beaucoup semblaient satisfaits de cette représentation.

   D'autres, en revanche, cherchaient à tracer un histogramme mais étaient confrontés à des difficultés techniques. Vu que tracer un histogramme n'était pas chose facile avec la version du tableur utilisée, l'assistant fournissait une aide technique aux étudiants. Le principe était le suivant : si l'étudiant dessine (au brouillon) un type de graphique en identifiant ce que représentent les axes et ce qui doit y figurer, l'assistant aide à le tracer

à l'aide du logiciel. Ainsi, certains groupes sont passés d'une représentation en nuage de points à une représentation en histogramme.

Lors de la mise en commun des différentes solutions proposées par les différents groupes, la représentation à l'aide d'histogramme (voir figure 3.16, bas) était globalement reconnue comme plus efficace et plus précise pour analyser les différences entre les trois séries de données.

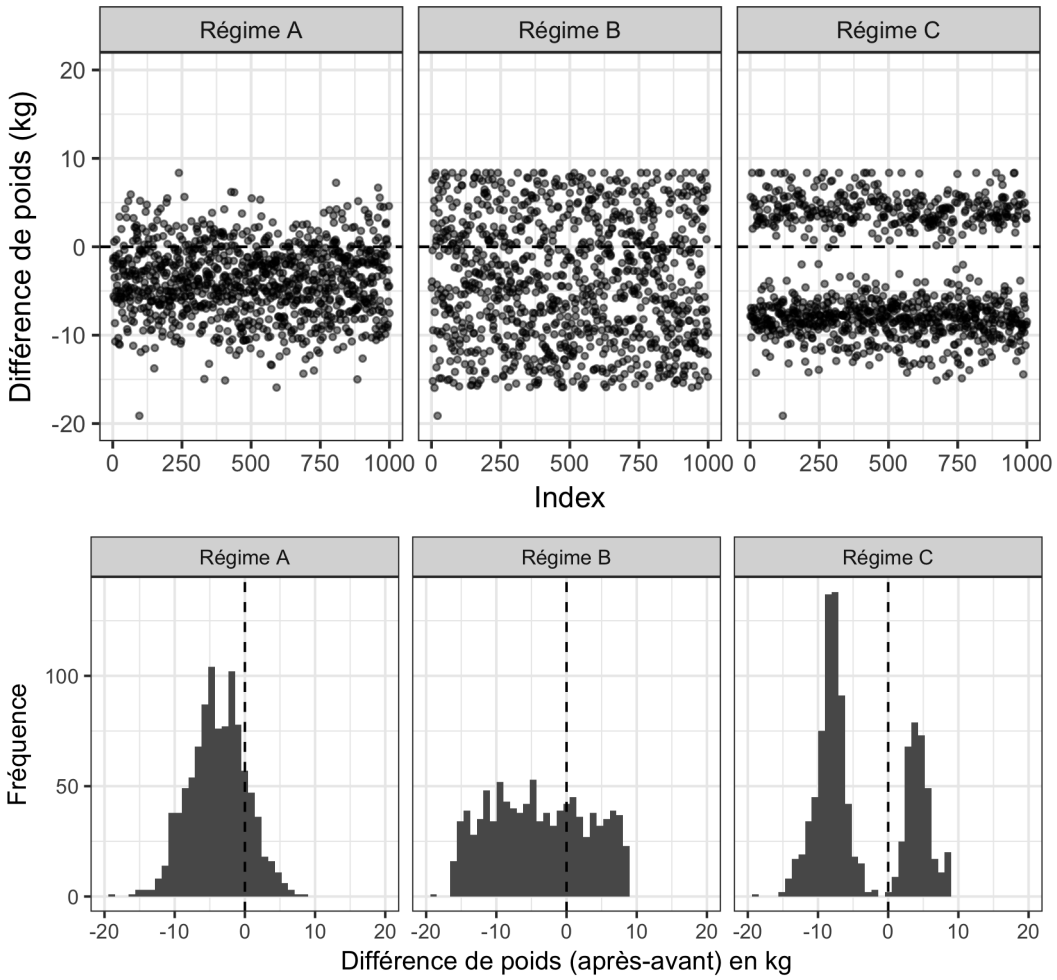


FIGURE 3.16 – **Représentation de la distribution des observations**. **Haut** : Représentation par les étudiants de la distribution des observations avec en ordonnées la perte de poids (kg) tandis que l'abscisse représente le numéro de la valeur dans la série, **Bas** : histogramme.

A côté de ces groupes d'étudiants qui ont *joué le jeu*, c'est-à-dire accepté de tenter de résoudre le problème à partir d'une réflexion basée sur les données disponibles et sur leurs conceptions, d'autres groupes ont essayé d'appliquer des formules mathématiques retrouvées dans le formulaire. Pour ces étudiants, la nature de l'exercice a donc changé puisqu'il s'agissait de trouver une justification aux formules choisies et appliquées : pourquoi calculer la variance et l'écart-type ? lequel choisir et pourquoi ? Qu'apportent-ils ?. Bien qu'intéressant en soi, l'exercice de justification se différencie de celui attendu puisque, d'une certaine manière, l'étudiant part

du principe qu'au moins une des formules qu'il a sous les yeux fournira une réponse adéquate au problème posé. La question est donc de savoir, parmi cette liste, quelles formules semblent convenir ou apporter une réponse satisfaisante.

Cette manière d'aborder le problème induit donc un certain appauvrissement de l'exercice proposé, cependant celle-ci nous semble assez difficile à éviter dès lors que l'on souhaite laisser une certaine liberté aux étudiants dans le choix de l'approche à mettre en œuvre pour résoudre le problème.

La phase d'institutionnalisation a été généralement assez riche en discussions, a permis de formaliser les notions visées en s'appuyant sur les connaissances construites par les étudiants. Elle a parfois permis d'entamer un débat sur la question de la manipulation des données : peut-on faire dire ce que l'on veut aux chiffres ?

Cette séance a aussi été généralement bien accueillie par l'équipe d'assistants qui se sont retrouvés autant à l'aise avec les méthodes pédagogiques (puisque celles-ci étaient bien acceptées par les étudiants) qu'avec les notions enseignées.

3.5.2 Séance 2

Réaliser une analyse *a posteriori* de la deuxième séance du dispositif est un exercice complexe car, comme expliqué précédemment, cette séance n'a pas été conçue ni modifiée selon les principes de la théorie des situations didactiques. En raison des contraintes qui pèsent sur le dispositif expérimental, en particulier le fait de devoir être mis en œuvre par une équipe d'assistants demandeuse d'une certaine stabilité dans le dispositif d'une année à l'autre, nous avions fait le choix de conserver cette séance telle quelle. Les conceptions didactiques implicites qui ont conduit à la conception de cette séance sont donc tout à fait différentes de celles qui furent à la base des séances 1 et 3. La séance s'approche plus d'un enseignement classique.

Il n'y a pas eu de dévolution du problème aux étudiants car ce n'était pas visé par cette séance. Ce qui était attendu des étudiants est très cadré par le fascicule qui accompagne la séance. Cela n'a pas semblé poser de difficultés car, dans l'ensemble, le contrat didactique a été bien accepté par les étudiants comme par les assistants, qui sont entrés dans le rôle qui leur était attribué.

Les connaissances qui auraient émergé durant la séance sont difficile à identifier car, à nouveau, cette séance n'avait pas pour objectif d'en faire émerger de nouvelles mais plutôt de présenter des notions et de demander aux étudiants de s'exercer à les utiliser ensuite.

Dans ce fonctionnement présentation-exercice, les notions engagées (manipulation de la distribution binomiale par exemple), n'ont pas semblé poser de difficultés majeures.

De même, les étudiants ont facilement répondu aux questions censées "provoquer une réflexion sur" l'écart entre une distribution modèle et une distribution observée. L'idée que la distribution simulée approche de la distribution théorique à mesure que le nombre d'observations est grand semble évidente au cours de cette séance et la notion de modèle ne semble pas poser de problème majeur.

On sait donc quels sont les comportements attendus, ceux-ci sont largement suggérés par l'assistant et par la structure de la séance, mais rien ne permet de dire que les étudiants ont réellement engagé leurs conceptions dans la tâche.

Globalement, on peut dire que le déroulement de la séance a été sans accrocs majeurs, mais que celle-ci n'était vraisemblablement pas de nature à remettre en cause les conceptions initiales des étudiants, ce qui peut expliquer pourquoi cette séance ne leur a pas posé de grands problèmes.

3.5.3 Séance 3

Un constat général

Quand on analyse les comportements des étudiants lors de la séance de travaux pratiques consacrée au test d'hypothèse, on constate qu'ils présentent de grandes difficultés à démarrer une réflexion qui aille dans la voie souhaitée par l'enseignant, c'est-à-dire qui permette d'aboutir à une réponse au problème posé.

Concrètement, cela se traduit par un blocage de la part des étudiants après que ceux-ci ont lu les consignes et au moment où ils cherchent une manière d'aborder le problème.

Ensuite, ceux-ci retournent vers l'enseignant (ou, pour les groupes enregistrés, l'investigateur) et celui-ci se sent dans l'obligation, au bout d'un certain temps, de reprendre la main et d'orienter les étudiants. On peut y voir l'effet du contrat didactique car l'enseignant se sent responsable de la réussite de l'activité d'enseignement. Bien que l'étudiant doive, au départ, jouer le jeu et tenter de résoudre le problème par lui-même, c'est l'enseignant qui est, *in fine*, responsable du fait qu'il y parvienne. C'est, en effet, lui qui a placé l'étudiant face au problème avec la garantie, implicite, que celui-ci était en mesure d'en venir à bout.

A partir du moment où l'enseignant accepte la responsabilité de la résolution du problème, les étudiants s'en remettent à lui pour évaluer leurs propositions. Dès ce moment, l'objectif des étudiants n'est plus tant de trouver une réponse au problème initial mais plutôt de trouver la réponse que l'enseignant attend. Et ils se tournent vers lui pour savoir si l'exercice est terminé, si "c'est bon comme cela" ou s'il faut aller plus loin. Ils ne se sentent, en effet, pas en mesure de juger du caractère abouti ou non de leur raisonnement.

On peut donc faire le constat que la dévolution du problème n'a, semble-t-il, pas eu lieu. La situation qui se voulait a-didactique et fondamentale ne l'a pas été.

Qui plus est, ce blocage a été récurrent : il s'est manifesté dans tous les groupes, c'est-à-dire chez près de 200 étudiants de cette cohorte. Il est donc difficile d'expliquer celui-ci par la mauvaise volonté de quelques individus.

En résumé, pour l'enseignant cette séance n'est pas satisfaisante car elle n'a pas permis de faire émerger, comme attendu, des connaissances qui auraient pu servir de base à un enseignement du test d'hypothèses. Les étudiants butent sur l'énoncé et ne parviennent à un début de solution qu'au prix de l'acceptation d'une logique qui leur est imposée.

Pour le chercheur, en revanche, les conclusions sont différentes. Certes, les choses ne se sont visiblement pas déroulées comme prévu mais ce qui compte c'est de voir ce que l'on peut apprendre de cette expérience. Certains éléments concernent la mise en œuvre du dispositif d'enseignement et d'autres concernent, plus fondamentalement, sa conception.

En ce qui concerne la mise en œuvre du dispositif d'enseignement, l'analyse des échanges entre les étudiants et l'investigateur révèle que les interventions de ce dernier ne sont pas toutes pertinentes. Les difficultés, blocages et interrogations des étudiants ont rendu nécessaires de telles interventions, d'autant qu'une partie de ces difficultés semblaient causées par la complexité superflue et artificielle liée au contexte. Ce n'est donc pas le fait que l'investigateur soit intervenu durant l'expérimentation qui nous semble poser problème mais plutôt la nature de ses interventions. En effet, à plusieurs reprises, en voulant se rattacher à des logiques potentiellement plus parlantes pour les étudiants, l'investigateur a emprunté des termes soit au test de significativité selon Fisher (la logique de la corroboration d'une hypothèse par les observations) soit aux intervalles de confiance (la logique de la précision et de la marge d'erreur). Ce faisant, il participe au maintien d'une certaine confusion entre ces différentes approches et rend la logique visée (celle du choix dans le test d'hypothèses) plus difficile à saisir.

Par ailleurs, l'analyse *a posteriori* nous permet d'identifier cinq éléments non anticipés ou insuffisamment pris en compte dans la conception de ce dispositif d'enseignement.

Le contexte

Le premier élément concerne le contexte, l'emballage médical autour de la tâche demandée aux étudiants. Les mises au point effectuées sur une précédente cohorte d'étudiants (voir analyse *a priori*) nous ont amené à ajouter un contexte médical à la tâche proposée devant les réactions dubitatives de certains étudiants envers la simplicité ou le caractère enfantin du matériel utilisé pour la situation (une boîte en carton et des blocs de couleur). Nous avons considéré qu'insérer le problème dans un contexte médical serait de nature à soutenir l'engagement des étudiants

dans la résolution de l'exercice et, au final, de mieux permettre la mise à l'épreuve de leurs conceptions.

L'analyse des étudiants en situation révèle plutôt que ceux-ci ne sont pas dupes et comprennent assez bien le caractère artificiel du contexte (voir Séance 3, groupe 2, point 1). De plus, celui-ci contient, en lui-même, une série d'éléments qui pourraient être intéressants dans d'autres domaines de la statistique ou pour d'autres disciplines et ne sont pas considérés pertinents ici.

Par exemple, dans le contexte, la question de l'efficacité vaccinale est abordée. Celle-ci serait différente selon les sous-types de virus. Certains étudiants font des remarques tout à fait pertinentes à propos de l'efficacité vaccinale (voir Séance 3, groupe 2, point 2). Et la question de la répartition des syndrômes grippaux en fonction de caractéristiques des individus (tels que le statut vaccinal) serait une question particulièrement intéressante dans le cadre d'un cours d'épidémiologie.

L'enseignant de statistique qui prépare une situation pour le test d'hypothèses voit aisément quels sont les éléments importants dans l'énoncé et quels sont ceux qui, en revanche, peuvent et doivent être négligés. Mais pour l'étudiant, qui *a priori* n'a pas le même bagage concernant l'inférence statistique, il n'est pas évident de savoir quelles difficultés sont dignes d'intérêt et lesquelles peuvent être simplement ignorées.

Le contexte utilisé dans ce dispositif est donc artificiel, complexe et il ne permet pas d'éclairer les choix à poser dans la résolution du problème qui est posé. En cela on pourrait le qualifier de pseudo-contexte.

La consigne

Une deuxième caractéristique de la situation et qui explique en partie la non-dévolution du problème aux étudiants est liée à la consigne principale : "Combien de prélèvements faudrait-il analyser pour connaître la proportion de virus A H3N2 durant l'épidémie cette année ? 10 ? 20 ? 50 ? 100 ou 200 ?"

Cette question nous avait semblé pertinente *a priori* car il s'agit de la question à laquelle les étudiants devront répondre en fin d'exercice. Ils devront, en effet, justifier un nombre de prélèvements à effectuer à partir du risque d'erreur que celui-ci impliquerait. Cependant, l'expérience montre que cette question arrive trop précocement et qu'elle ne permet pas de lancer les étudiants dans une réflexion qui doit aboutir *in fine* à la détermination d'une taille d'échantillon.

A ce titre, il aurait sans doute été intéressant de procéder par étapes à l'instar de ce que Brousseau (2005) et Régnier (1998) ont fait. Ainsi, une question de départ aurait pu être sim-

plement : "Quel est le contenu de boîte ?" ce qui aurait amené les étudiants à réaliser des tirages et à proposer une réponse. Celle-ci aurait pu alors servir de base à d'autres questions, du type : "A partir de quand peut-on affirmer connaître le contenu de la boîte ?" pour, progressivement, arriver à une question comme "Combien de tirages faudrait-il pour connaître le contenu de la boîte ?".

Le type de raisonnement attendu

Le troisième élément qui nous semble expliquer les écarts entre les comportements attendus et les comportements observés est l'important décalage entre les attentes de l'investigateur et celles des étudiants concernant le type de démarche à mettre en œuvre dans cet situation.

L'analyse sémantique soutient l'idée que l'investigateur essaie d'orienter les étudiants vers une démarche de *quantification* d'un degré de *compatibilité* entre des *résultats obtenus* et des *hypothèses*. Il insiste également sur la notion de *risque* et sur le fait que cette situation est une *simplification* d'une réalité plus complexe. De leur côté, les étudiants cherchent à comprendre ce qu'ils *doivent* faire et s'attendent à devoir entrer des *nombres* dans une *formule* impliquant une distribution théorique telle que la distribution *binomiale*, et ce en suivant une procédure vue au *cours*.

C'est, en substance, ce qu'admet une étudiante en fin de séance :

E8 (groupe 4) : "*Maintenant je ne vais pas le cacher, ce n'est pas du tout comme ça que je pensais qu'on allait (résoudre le problème) (...) Et mais même, jamais je n'aurais eu l'idée d'aller faire ça ! Moi, pour moi, une petite formule ça allait être (suffisant). Taper mon chiffre, mes possibilités et (appliquer la formule et avoir la réponse).*"

Les étudiants croient reconnaitre un exercice d'analyse combinatoire et tentent d'agencer les différentes données disponibles (5 boules, de deux couleurs différentes, 10 tirages par exemple) à l'aide de raisonnements mis en œuvre dans des cours d'analyse combinatoire ou de probabilités qu'ils ont pu avoir par le passé (en dernière année de secondaire ou simplement lors de la dernière séance de travaux pratiques de statistique)...

E1 (groupe 1) : "*On l'a fait la dernière fois avec le tirage de pièces. (...) J'ai vu une loi l'année dernière où on prenait des chiffres de tel à tel nombre et qu'on pouvait les prendre plusieurs fois ou indépendamment mais je ne me souviens plus de la formule*".

E6 (groupe 3) : "*Ca servirait à quelque chose de faire toutes les combinaisons possibles et du coup de voir les probabilités de chacun ? C'est ce qu'on a fait au cours passé*".

E8 (groupe 4) : "(...) *je suis sûre qu'on a vu (quelque chose) dans le même genre en rhéto, l'analyse combinatoire.*"

. . . avec parfois des tentatives de calculs dépourvus de sens qui ne sont pas sans rappeler la célèbre expérience de "l'âge du capitaine", c'est-à-dire que les étudiants combinent les données au sein d'un calcul qui n'a aucun sens.

E2 (groupe 1) : "*J'ai calculé (que) pour 10 (?) ça fait 192. Mais en faisant ça je ne sais pas comment on va trouver*".

Ou encore :

E4 : (groupe 2) "*On peut avoir une marge d'erreur à 20 % près... donc 50 (tirages) alors ? (Pourquoi ?) Je ne sais pas, j'ai fait un rapport 5 boules fois 10 (rires).*"

Pour les étudiants, nombreux, qui ont l'idée, et parfois la certitude, qu'il s'agit d'un problème d'analyse combinatoire comme ils en ont déjà eu, il est difficile d'envisager de mettre en œuvre un raisonnement autre.

Il est, finalement, assez logique [20] qu'ils ne se dirigent pas vers un problème d'inférence car il s'agit, pour la plupart des étudiants, de leur première confrontation avec ce type de problème. Aux yeux de ces étudiants, l'analyse combinatoire est le domaine qui s'approche le plus de ce qui leur est demandé ici.

E6 : "*Ça me paraît bien dans le sens où il faut déterminer N, c'est un des seul trucs que je connais où il y a N.*"

On peut expliquer décalage entre les attentes de l'investigateur, d'une part, et des étudiants, d'autre part, par un effet de contrat didactique. Les attentes des étudiants concernant la manière dont ils allaient résoudre le problème sont probablement légitimes au regard de ce qui leur a été proposé jusqu'à présent, lors des précédentes séances de travaux pratiques ou des autres cours. Dans ce contexte, la manière dont la situation a été présentée aux étudiant n'a pas permis d'opérer une rupture suffisamment claire avec ces attentes.

Le statut de l'échantillon

Un quatrième élément non anticipé et qu'il nous semble important de prendre en compte à l'avenir est l'existence d'une conception chez les étudiants, qui concerne le statut de l'échantillon et qui pourrait bien faire obstacle au développement de connaissances concernant l'inférence statistique.

Cette conception s'est assez clairement manifestée lors de l'évaluation écrite qui a eut lieu en fin d'année académique (2018-2019) dans le cadre du cours de biostatistique avec le même public que celui qui a participé aux quatre séances de travaux pratiques précédemment décrites.

En substance cette conception (C1, ci-après) peut s'énoncer de la manière suivante : l'échan-

20. Surtout *a posteriori*.

tillon sert à dire quelque chose à propos d'une population qui est déjà connue par ailleurs. Un échantillon doit donc être choisi pour être représentatif *individuellement* de la population cible. Les outils d'inférence statistiques permettent d'évaluer le caractère représentatif ou non d'un échantillon.

Nous allons voir que cette conception :

1. Possède probablement un domaine de validité important chez ces étudiants ;
2. Diffère sensiblement du raisonnement statistique ;
3. Est de nature à expliquer différentes erreurs, récurrentes chez ces étudiants.

Intéressons-nous d'abord au **domaine de validité** de cette conception. Qu'est-ce qui pourrait expliquer que celle-ci soit présente parmi les étudiants et efficace dans un certain nombre de situations ?

Il semble que cette conception possède un domaine de validité particulièrement étendu, en particulier parmi des étudiants du domaine bio-médical.

Prenons quelque exemples.

Quand ces étudiants dessinent une mitochondrie, sur base d'images issues d'un microscope électronique par exemple, ils observent un grand nombre de mitochondries (assimilable à une population) et ils en choisissent une qui leur semble suffisamment représentative de la population pour la dessiner. C'est également ce que fait le professeur au cours.

Quand, au cours d'anatomie, le professeur représente, par exemple, le système digestif, il ne montre pas le système digestif d'un individu tiré aléatoirement parmi une certaine population mais plutôt un système qui lui semble suffisamment représentatif de la population, population qui ne lui est pas tout à fait inconnue.

Quand un chercheur montre les effets du tabagisme sur la structure d'un poumon, ou les effets d'un polluant sur la morphologie de poissons, il ne sélectionne pas au hasard l'image qu'il expose, il en prend une qui lui semble représentative (ou suffisamment convaincante).

Dans chacun de ces exemples, la population est globalement connue dès le départ ou, du moins, on considère[21] qu'on la connait : soit parce que l'on a une vue sur des milliers de mitochondries sur une coupe de microscope, soit parce que l'on a déjà une idée de ce que devrait être l'anatomie du système digestif ou encore parce que l'on est convaincu de l'effet néfaste du polluant. Puisque la population est globalement connue, le rôle de l'échantillon sera d'en être un bon représentant. On sélectionnera donc un échantillon qui soit bien orienté, net, sans artefact et qui corresponde à l'image que l'on se fait de la population. On pourrait multiplier les exemples de disciplines dans lesquelles un tel raisonnement est valide : biologie cellulaire,

21. Peut-être à tort.

zoologie, microbiologie, anatomie, imagerie médicale, botanique, histologie, physiologie, endocrinologie, immunologie *etc.*

La conception C1 semble également trouver quelque support en dehors de la sphère scolaire de ces étudiants.

Ne dit-on pas dans les médias "Cette étude a été réalisée sur un échantillon *représentatif* de 1000 personnes ?". La formulation de cette expression est tout à fait en adéquation avec l'idée que l'échantillon doit être représentatif d'une population qui est globalement connue. Pourtant, en réalité, si on prend le cas d'un sondage d'opinion politique, le caractère représentatif ou non de l'échantillon concernera les caractéristiques *démographiques* de l'échantillon (âge, sexe, région, *etc.*) qui sont, en effet, bien connues au niveau de la population mais il ne concernera pas *les opinions politiques* qui, par définition ne sont pas connues puisque l'on fait un sondage pour s'en faire une idée. Pour éviter la confusion, il serait peut-être bon de parler d'un échantillon qui est représentatif de la population en ce qui concerne les caractéristiques démographiques (mais dont on ignore s'il est, au niveau des opinions politiques, proche ou éloigné de ce qu'on obtiendrait si on posait la question à l'ensemble de la population).

Le concept "d'échantillon gratuit" que l'on peut retrouver dans certains magasins supporte également la conception C1 puisque cet "échantillon" représente individuellement la "population" (le flacon de parfum entier, par exemple) qui est, par ailleurs, connue. Le lien entre le concept statistique d'échantillon et l'échantillon que l'on retrouve dans les magasins a d'ailleurs été retrouvé chez des étudiants universitaires et a été montré dans l'expérience de Régnier (1998).

Bref, la conception C1 pourrait, tout à fait, avoir un assez grand domaine de validité que ce soit dans la sphère scolaire ou en dehors de celle-ci.

Pourtant, dans le champ de la statistique, cette conception va totalement à l'encontre du **raisonnement statistique**.

Pour le statisticien, en effet, la conception correcte serait plutôt la suivante : "L'échantillon, s'il a été généré par un procédé aléatoire, est représentatif *en moyenne* de la population qui, elle, est inconnue par définition".

Individuellement, un échantillon peut donc s'écarter de la moyenne de la population mais tant que celui-ci est issu d'un tirage aléatoire, on sait qu'en moyenne, l'estimation qu'il fournira sera correcte.

Si l'échantillon ne provient pas d'un tirage aléatoire au sein de la population, alors il y un risque de *biais*, c'est-à-dire de décalage *en moyenne* entre les estimations fournies par les échantillons et la valeur cible au niveau de la population.

En cas de tirage aléatoire et indépendant, le statisticien sait que les caractéristiques indi-

TABLE 3.13 – Corollaires de la conceptions C1

Bon échantillon	**Mauvais échantillon**
non-biaisé	biaisé
représentatif	non-représentatif
fiable	non-fiable
P-valeur < 5 %	P-valeur > 5 %
estimation comprise dans l'intervalle de confiance	estimation non comprise dans l'intervalle de confiance
rejet de l'hypothèse nulle	acceptation de l'hypothèse nulle
n suffisant	n insuffisant

viduelles de l'échantillon s'écarteront plus ou moins de celles de la population et donc il porte une attention toute particulière à l'étendue de cet écart, à la notion de *variabilité*.

Dans ce cadre, le statisticien utilise différents outils d'inférence statistique pour se prononcer à propos de la population inconnue : outils d'estimation ou de test.

La conception C1 se différencie donc assez nettement de la conception que l'on pourrait identifier comme correcte dans le champ de la statistique et elle semble y déformer certaines notions statistiques telles que la population, le biais, l'intervalle de confiance ou encore la P-valeur. Dans le champ de la statistique, la conception C1 semble donc impliquer les **corollaires** suivants :

— la population est connue mais trop grande pour être décrite ;
— la population est bien décrite par l'échantillon pourvu que celui-ci soit bon, représentatif (voir tableau 3.13) ;
— si l'échantillon ne représente pas bien la population, il y a un biais ;
— l'intervalle de confiance sert à vérifier que l'échantillon est bien représentatif de la population (si la valeur estimée se trouve incluse dans l'intervalle) ;
— un échantillon est représentatif si la P-valeur < 5 % ou si le test d'hypothèses conduit à un rejet d'hypothèse nulle.

Au vu de l'éloignement entre la conception correcte et la conception C1, on pourrait penser que cette dernière sera vite reconnue comme inefficace par les étudiants qui suivent le cours de statistique. Cependant, notre expérience d'enseignement nous suggère qu'il en va autrement et que ce genre de conception (C1) explique assez bien les trois **erreurs récurrentes** que voici :

1. un échantillon trop petit risque de biaiser les résultats d'un essai clinique car il risque de ne pas être représentatif de la population ;

   Lors de l'évaluation certificative (juin 2019) de la cohorte d'étudiants en médecine sur

laquelle le dispositif expérimental a été testé (printemps 2019), une des questions était la suivante : "Le fait de conduire un essai clinique sur un échantillon trop petit peut-il biaiser les résultats d'un essai clinique ?". La réponse attendue était que non, il en résultera une plus grande imprécision mais il n'y a pas de raisons que cela provoque un biais. Or, nous avons été surpris de constater que 90 % des étudiants ont répondu que oui. La justification la plus courante était qu'un petit échantillon risquait de ne pas être représentatif de la population.

2. la P-valeur mesure la fiabilité d'un résultat, quand elle est inférieure au seuil de 5 % cela indique que l'échantillon est fiable et qu'il peut être utilisé pour représenter la population ;

   Ce type d'erreur revient de manière récurrente chez les étudiants et se retrouve encore dans les réponses de chercheurs.

3. l'intervalle de confiance permet également de déterminer si l'échantillon choisi est *bon* ou *mauvais* selon que la valeur estimée se retrouve ou non entre les bornes de l'intervalle.

   Ce type d'erreur est également extrêmement récurent dans les productions d'étudiants.

Selon nous, ce genre d'erreurs pourrait assez bien s'expliquer par une conception sous-jacente proche de la conception C1 décrite plus haut. Nous pensons donc que la conception C1 est potentiellement un obstacle à l'apprentissage de l'inférence statistique de manière générale et devrait être pris en compte dans l'élaboration de dispositifs d'enseignement de l'inférence statistique.

La relation entre la taille d'échantillon et le risque d'erreur

Notons également que les échanges entre les étudiants sont cohérents avec une certaine conception concernant la relation entre la taille de l'échantillon et le risque d'erreur.

On y trouve une conception de la variabilité ne parvenant pas à dépasser le niveau qualitatif.

E1 : "*Bin on aura en grande majorité ceux-là et ça c'est difficile à obtenir*"

I : "*Oui, ça c'est au niveau qualitatif (. . .) mais est-ce qu'on ne sait pas le mesurer ?*"

E2 : "*C'est un truc de probabilités mais je ne sais pas c'est lequel*".

Ou encore :

E3 : "*C'est pas impossible de se retrouver avec que des H1N1 dans les 10 personnes, mais en soi, même avec 100 c'est pas impossible, ça peut toujours arriver qu'on se retrouve que avec des H1N1*".

I : "*Et on ne sait pas le mesurer ?*"

E4 : "*Mais oui mais pour le mesurer il faudrait des valeurs*".

On peut également citer l'idée qu'une faible taille d'échantillon ne permettrait pas des inférences suffisamment fiables.

I : "*Pourquoi (n=10) c'est pas assez ?*"

E5 : "*Parce qu'on ne peut pas généraliser une proportion de tous les malades avec un échantillon (aussi faible)*".

E4 : "*On pourrait tomber que sur des gens qui ont l'autre grippe ou soit que sur des gens qui ont le H3N2*".

Enfin, on notera également que les raisonnements du groupe 2 sont cohérents avec une idée de stabilité des résultats à partir d'une certaine taille d'échantillon.

I : "*Mais ici vous l'avez fait une fois (n=50). Cette fois-là ça s'est stabilisé mais peut-être que si on le refaisait, il faudrait 100 lancers pour que ça se stabilise, non ?*"

E5 : "*Non parce que franchement ça se voit que c'est stable*"

On peut interpréter les comportements des étudiants comme étant révélateurs de la présence d'une certaine conception concernant la variabilité (ci-après conception C2) : l'étudiant admet l'existence d'une certaine variabilité, donc d'un certain risque d'erreur, mais semble agir comme si cette variabilité disparaissait au-delà d'une certaine taille d'échantillon.

Pour décrire la relation entre taille d'échantillon et risque d'erreur, donnons une définition plus précise de l'expression *risque d'erreur*.

Dans le contexte étudié, les différentes hypothèses statistiques sont espacées de 20 % ($\pi = 0$, $\pi = 0,2$, $\pi = 0,4$, $\pi = 0,6$, $\pi = 0,8$ et $\pi = 1$), appelons cet écart, δ. Si l'on considère une seule hypothèse (par exemple celle selon laquelle $\pi = 0,6$) et que l'on choisit de considérer que cette hypothèse est correcte dès lors que la proportion observée tombe dans l'intervalle $]\pi - \frac{\delta}{2}; \pi - \frac{\delta}{2}]$ [22], alors le risque d'erreur peut se définir comme la probabilité, sous cette hypothèse, d'obtenir une proportion tombant en dehors de l'intervalle $]\pi - \frac{\delta}{2}; \pi - \frac{\delta}{2}]$.

Dans ce cas, la figure 3.17 (gauche) représenterait la conception C2 concernant l'évolution du risque d'erreur en fonction de la taille d'échantillon.

On y trouve l'idée qu'en dessous d'une certaine valeur, les données ne sont pas fiables, que la variabilité est trop grande, que le risque d'erreur n'est pas négligeable et que celui-ci le devient à partir d'une certaine taille d'échantillon. La relation ainsi décrite n'est pas précise ce qui est en phase avec le fait que les étudiants ne semblent pas encore maitriser d'outil permettant une prise en compte quantitative de la variabilité des résultats aux différentes tailles d'échantillon.

22. Voir méthode des intervalles fixes dans les raisonnements possibles décrite dans l'analyse *a priori*.

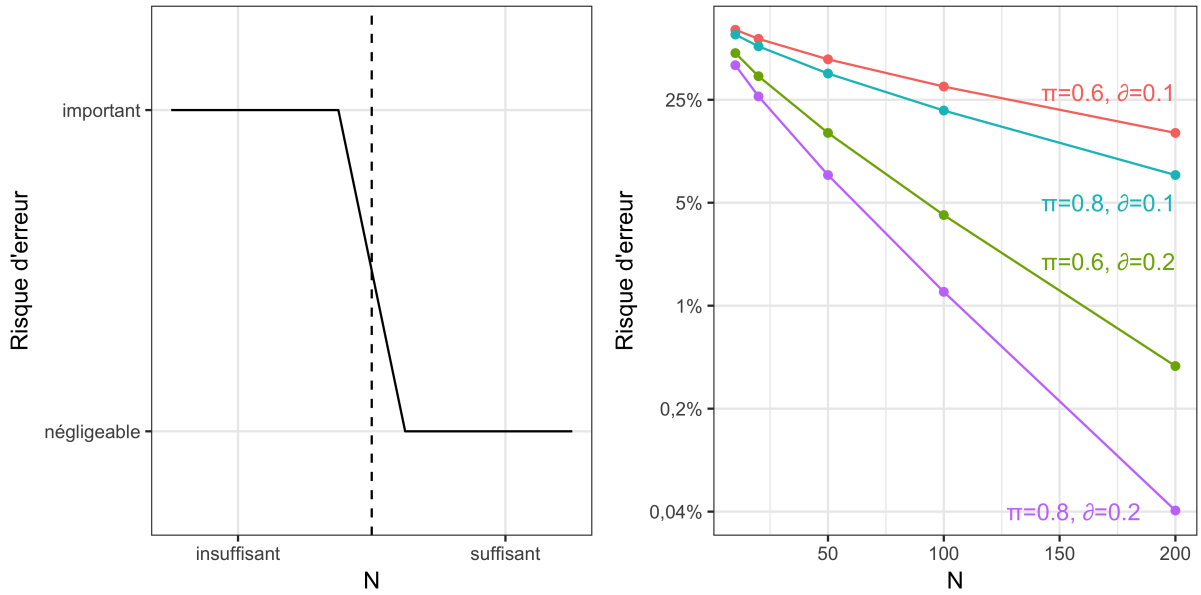


FIGURE 3.17 – **Deux manières de concevoir la relation entre le risque d'erreur et la taille de d'échantillon**. A gauche, une conception "simple" (C2), à droite une conception plus complexe de cette relation. Les risques d'erreurs ont été calculés à partir de la loi normale.

Une conception plus éclairée de la relation entre taille d'échantillon et risque d'erreur est représentée sur la figure 3.17 (droite).

La relation diffère par plusieurs aspects. D'une part, elle est définie quantitativement et plus uniquement qualitativement. D'autre part, elle dépend de deux variables : l'écart entre les hypothèses à départager (δ) – le risque d'erreur sera d'autant plus important que les hypothèses à départager sont proches (et que δ sera petit) – et la variabilité entre les individus (fonction de π [23]) – le risque d'erreur étant d'autant plus important que les individus sont variables.

On peut noter que, dans l'expérience que nous décrivons, la conception C2 n'a pas été suffisamment éprouvée, ne permettant pas aux étudiants d'en apercevoir les limites. Pour y parvenir, il aurait sans doute fallu jouer sur les variables didactiques suivantes :

1. Le nombre de réalisations de l'expérience : amener les étudiants à recommencer leurs tirages devrait permettre d'engager une réflexion sur la quantification de la variabilité des résultats.

2. Le couple (δ, π) : se placer dans des situations impliquant un besoin de précision plus ou moins grand (en faisant varier δ) et une variabilité plus ou moins importante (en faisant varier π) devrait permettre de complexifier progressivement la conception C2 pour prendre en compte les aspects liés à la précision et à la variabilité dans la relation entre taille d'échantillon et risque d'erreur.

23. Dans le cas où les observations se résument à l'aide d'une proportion.

3.5.4 Séance 4

La quatrième séance était censée voir les étudiants présenter et défendre leur démarche de résolution du problème puis l'enseignant formaliser les concepts émergents au cours de la phase d'institutionnalisation.

Cependant, nous avons vu, dans l'analyse de la troisième séance de travaux pratiques, que, pour différentes raisons, la dévolution du problème n'a pas été observée. L'enseignant ou l'investigateur ont, chaque fois, dû reprendre la responsabilité de la résolution du problème et les étudiants se sont donc reposés sur celui-ci pour valider leur raisonnement.

Il en résulte que la quatrième séance n'a pas été une phase de formalisation des conceptions qui auraient émergé chez les étudiants confrontés à la situation fondamentale du test d'hypothèses. En effet, la démarche que les étudiants ont présentée lors de cette séance n'étaient pas réellement la leur mais plutôt celle suggérée par l'enseignant ou l'investigateur. L'exercice était donc plutôt un exercice de restitution d'une démarche qui leur a été soufflée plutôt qu'un exercice de justification d'une démarche qu'ils auraient fortement contribué à construire.

L'institutionnalisation réalisée par l'enseignant a donc porté sur les notions restituées par les étudiants plutôt que sur des notions construites par eux.

3.6 Conclusions

La question de recherche que nous avons posée en début de chapitre était la suivante.

Dans quelle mesure le recours à une situation conçue pour reproduire les conditions d'émergence du test d'hypothèses permet-il aux étudiants des filières biomédicales à l'Université de Namur d'en apprendre la logique sous-jacente ?

A partir de cette question générale, nous avons formulé les questions spécifiques suivantes.

A quel point la situation mise au point permet-elle d'atteindre les objectifs visés ? Permet-elle de faire émerger des connaissances pouvant servir de base à un apprentissage satisfaisant de la logique sous-tendant le test d'hypothèses ? Si oui, dans quelle mesure ? Et si non, quelles conceptions font obstacle à cet apprentissage ?

L'analyse *a posteriori* nous permet d'apporter les réponses suivantes.

Telle qu'elle a été conçue et mise en œuvre, le recours à une situation conçue pour reproduire les conditions d'émergence du test d'hypothèses ne constitue pas un point de départ adéquat pour l'apprentissage de la logique sous-jacente au test d'hypothèses.

En effet, nous observons un blocage de la part des étudiants confrontés à cette situation et un échec de la dévolution, que ce soit dans les quatre groupes d'étudiants dont les verbatim ont été analysés ou dans le reste de la cohorte d'étudiants. La situation n'a pas permis, chez ces étudiants, l'émergence de connaissances sur lesquelles baser un enseignement satisfaisant de la logique du test d'hypothèses. La situation ne peut donc pas être qualifiée de fondamentale car elle n'est pas adidactique et ne parvient pas à rendre les conceptions visées plus efficientes que les conceptions initiales.

Si les observations faites au cours de cette expérience ne permettent pas de valider l'hypothèse initiale, elle permettent de dégager des pistes pour l'enseignement de l'inférence statistique. Ces pistes concernent les dispositifs d'enseignement, d'une part, et, d'autre part, les conceptions des étudiants.

Concernant **le dispositif d'enseignement**, l'analyse de la première séance de travaux pratiques, consacrée aux notions de statistique descriptive, montre qu'une approche basée sur une situation adidactique conçue pour favoriser l'émergence de certaines notions en statistique est tout à fait en mesure de produire les effets attendus, c'est-à-dire faire émerger, chez ces étudiants, des connaissances à la fois nouvelles et proches des notions visées qui permettent, dans un deuxième temps, une institutionnalisation des connaissances ainsi construites.

Si l'on compare les caractéristiques des situations utilisée pour l'analyse descriptive (séance 1) et pour le test d'hypothèse (séance 3), on peut noter au moins trois différences importantes.

Premièrement, dans la séance 1, le contexte est utile et a du sens. Le fait de travailler sur des mesures de masses corporelle permet, par exemple, de savoir ce qui constitue des différences dignes d'intérêt (une différence de 2 kg, par exemple) et ce qui, au contraire, est plutôt négligeable (une différence de 50 g, par exemple). A l'opposé, le contexte de la séance 3 n'aide en rien à faire des choix dans la résolution du problème. Au contraire, ce contexte apporte un niveau de complexité supplémentaire dont les étudiants doivent se défaire avant de pouvoir réellement entamer leur réflexion.

Analyser des données implique nécessairement de poser des choix entre différentes voies possibles et cela n'est possible que dans un contexte qui a du sens et qui est un minimum maitrisé. Dans le cadre du test d'hypothèses, par exemple, il faut être en mesure de définir les hypothèses concurrentes et de discuter de la balance entre la taille d'échantillon et les risque d'erreurs.

Il apparait dès lors essentiel d'éviter à l'avenir les pseudo-contextes et d'*appliquer* les notions de statistique *appliquée* sur des données suffisamment authentiques.

Deuxièmement, la consigne initiale de la séance 1 "Comparer l'évolution du poids de patients soumis à trois régimes différents dans le but de déterminer celui qui semble le meilleur" est bien plus accessible à ces étudiants que la consigne de la séance 3 "Déterminer le nombre

de prélèvements à analyser pour connaitre la proportion de virus H3N2 durant l'épidémie de grippe de cette année". La première consigne permet rapidement aux étudiants de s'emparer du problème et de tester des solutions basée, dans un premier temps, sur leurs conceptions initiales. La consigne de la séance 3, au contraire, ne fait référence à rien de suffisamment connu pour ces étudiants. Elle touche à une question que les étudiants ne se poseraient pas spontanément et dont ils ne mesurent pas vraiment l'enjeu. Or si les étudiants ne comprennent pas vraiment le problème au départ, il leur est difficile d'imaginer des solutions.

Enfin, troisièmement, l'écart entre ce que les étudiants savent et les connaissances qu'ils sont censés construire parait bien plus adapté dans la séance 1 que dans la séance 3. Dans cette dernière, l'obstacle à franchir semble si important, le niveau auquel les étudiants doivent réorganiser leurs conceptions est si élevé, qu'il semble déraisonnable de penser qu'ils puissent y parvenir en si peu de temps.

Par ailleurs, l'analyse sémantique des concepts mobilisés par les étudiants et par l'investigateur montre à quel point les attentes des uns et des autres sont éloignées. Les étudiants ne trouvent pas la solution au problème mais, surtout, ils ne comprennent pas le type de solution attendu. Ils cherchent à appliquer une formule ce qui, d'une certaine manière, a fonctionné lors de la séance 1, là où l'investigateur essaye de les pousser à mettre en œuvre une démarche complexe impliquant la quantification d'un degré de compatibilité entre des observations et des hypothèses.

A l'avenir, il semble important, d'une part, de mieux prendre en compte les attentes des étudiants concernant le type de résolution attendu, et de clarifier le contrat didactique si nécessaire. D'autre part, il faudra prendre en compte le fait que l'écart entre les conceptions initiales des étudiants et les notions visées est bien plus important avec la logique sous-jacente au test d'hypothèses qu'avec les outils d'analyse descriptive. Aborder le test d'hypothèses au cours d'une séance de deux heures ne permet pas aux étudiants d'en intégrer véritablement la logique sous-jacente.

Concernant les **conceptions** des étudiants, l'analyse *a posteriori* permet également d'identifier, au travers des raisonnements des étudiants, au moins deux conceptions non-anticipées et pouvant faire obstacle à l'apprentissage de la logique sous-jacente au test d'hypothèse.

La première concerne le rôle de l'échantillonnage. On constate, en effet, que pour certains étudiants l'échantillonnage peut se voir comme une manière de décrire une population qui est déjà connue par ailleurs. L'échantillon doit, dès lors, constituer un représentant correct de la population, il doit être individuellement représentatif de celle-ci. Cette conception est cohérente avec des situations où l'échantillon sert, non pas à apprendre quelque chose d'une population d'intérêt inconnue, mais plutôt à illustrer une population connue mais trop grande que pour être décrite dans son intégralité. On peut imaginer que travailler dans des contextes où la population d'intérêt est réellement inconnue est indispensable pour mettre à mal cette conception.

La seconde concerne la relation entre risque d'erreur et taille d'échantillon. Les raisonnements d'étudiants semblent, en effet, assez compatibles avec l'idée que la variabilité existe mais ne pose problème, par exemple en termes de risque d'erreur, qu'en dessous d'une certaine taille d'échantillon. Cette conception pousse l'étudiant à analyser des résultats expérimentaux de manière binaire : il y a trop peu d'individus donc les résultats ne sont pas fiables ou, au contraire, il y a assez d'observations donc les résultats sont fiables. Or, dans le domaine de la recherche biomédicale, au vu de la variabilité inter-individuelle, de la faible intensité des effets étudiés et de la difficulté d'obtenir de très grands échantillons (par exemple au delà de 1000 individus), on se situe très souvent dans un cas de figure intermédiaire. Le risque d'erreur n'est pas négligeable mais les données doivent tout de même être exploitées. Dans ce contexte, il devient indispensable de dépasser la conception simple que les étudiants peuvent avoir et de les amener à en construire une plus complexe permettant, notamment, de quantifier les risques d'erreurs dans différents cas de figure. Pour y parvenir, il semble important de travailler à partir de contexte différents en termes de variabilité ou de précision nécessaire.

Ces deux conceptions devront être prise en compte dans nos réflexions visant à améliorer l'enseignement de l'inférence statistique. Il sera, en particulier, nécessaire de réfléchir au choix des variables didactiques afin de se placer dans des contextes à même de déstabiliser ces conceptions.

En conclusion, si le dispositif expérimental n'a pas permis de faire émerger chez les étudiants, les connaissances visées, l'analyse *a posteriori* que nous en faisons permet de dégager de nombreux d'éléments nouveaux ou insuffisamment pris en compte dans l'analyse *a priori* et constituant autant de pistes pour améliorer l'enseignement de l'inférence statistique chez ce public.

Chapitre 4

Discussion

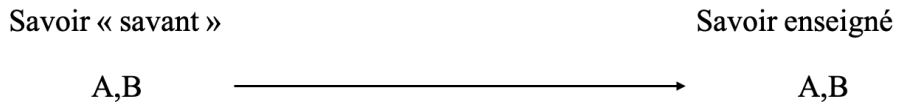
Dans ce dernier chapitre nous proposons, dans un premier temps, de tenter de répondre à la question soulevée dans l'introduction. Nous reviendrons, ensuite, sur les principaux choix sous-tendant le présent travail et discuterons de ses principales limites. Enfin, nous verrons quels peuvent-être les apports de ce travail et les perspectives qu'il dégage.

4.1 Synthèse

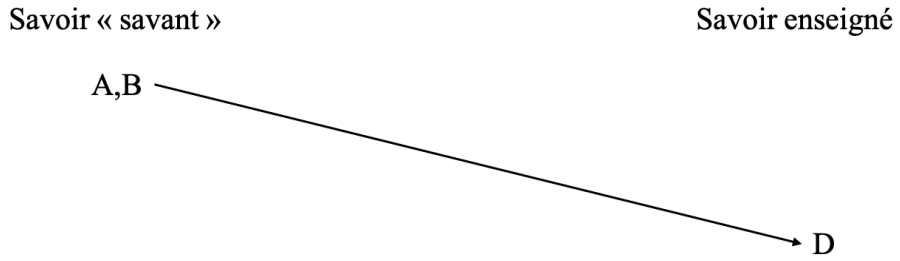
Dans cette section, nous verrons en quoi la réflexion menée jusqu'à présent apporte des éléments de réponse à la questions générale de la thèse : "*Comment améliorer l'enseignement de l'inférence statistique ?*".

Dans le chapitre 2, nous interrogeons la transposition des outils d'inférence statistique entre le savoir dit "savant" et le savoir enseigné localement. Cette analyse permet, premièrement, de montrer les écarts entre le savoir savant et le savoir enseigné, deuxièmement, de faire une hypothèse quand à la cause de ces écarts et, troisièmement, de suggérer des pistes pour améliorer l'enseignement.

Questionner le savoir enseigné ne va pas de soi car on pourrait penser que ce qui est enseigné à l'université correspond nécessairement à ce que l'on retrouverait au niveau du savoir savant. Certes, on peut attendre de légères différences ou simplification des concepts initiaux mais pas de quoi dénaturer ceux-ci. C'est ce que le schéma ci-dessous représente, A et B étant deux praxéologies existant au niveau du savoir savant et "dupliquées" (non transposées) dans le savoir enseigné.



Le premier constat que l'étude de la transposition didactique nous amène à poser est que, en ce qui concerne l'inférence statistique et au niveau local, le savoir enseigné s'écarte substantiellement de ce que l'on peut décrire au niveau du savoir savant [1]. Là où la praxéologie A (le test de significativité selon Fisher) vise à mesurer un degré auquel les observations corroborent une hypothèses d'intérêt et où la praxéologie B (le test d'hypothèses selon Neyman et Pearson) propose une méthode permettant d'opérer des choix entre plusieurs hypothèses concurrentes en contrôlant les risques d'erreur sur le long terme, la praxéologie D (le test d'hypothèses tel qu'enseigné) constitue un outil de mise en évidence d'effets expérimentaux en présence de variabilité expérimentale.



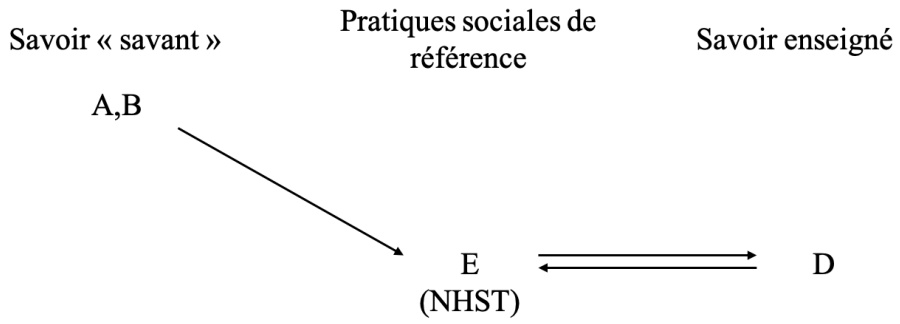
Pour comprendre l'ampleur de la transposition didactique à l'œuvre, il est nécessaire de s'intéresser aux pratiques sociales de référence, les pratiques d'analyses de données expérimentales dans la recherche biomédicale en l'occurrence. Les praxéologies mises en œuvre dans les pratiques professionnelles semblent, en effet, servir d'intermédiaire entre le savoir savant et le savoir enseigné.

L'examen des pratiques d'inférence statistique en recherche biomédicale révèle l'existence d'une praxéologie à la fois extrêmement répandue et vivement critiquée, le *Null Hypothesis Significance Testing* (NHST, praxéologie E).

1. Schématiquement nous représenterons cette différence de nature dans l'objectif des outils par un décalage vertical, les savoirs ne sont plus "alignés".

La spécificité du NHST par rapport aux praxéologies existant dans le savoir savant est de faire l'impasse sur la prise en compte, nécessairement subjective, des éléments de contexte. En effet, le test de significativité (A) nécessite d'énoncer l'hypothèse de recherche et d'interpréter la P-valeur en fonction du contexte. Dans le test d'hypothèses proposé Neyman et Pearson, l'expérimentateur détermine, au préalable et selon le contexte, les hypothèses à comparer et les risques d'erreurs considérés acceptables. A l'inverse, le NHST confronte les observations à une hypothèse générique – l'effet étudié n'existe pas – pour aboutir à une conclusion binaire : l'effet est démontré ou l'effet n'est pas démontré. Le NHST apparait dès lors comme une méthode universelle permettant de fournir une démonstration objective de l'existence d'effets expérimentaux.

Le NHST constitue une praxéologie distincte de ce qui est décrit dans le savoir savant mais assez proche de ce qui est enseigné. Si, dans le cas présent, les pratiques professionnelles influencent indubitablement le savoir enseigné, l'inverse est également vrai car les chercheurs, dans une certaine mesure, appliquent ce qui leur a été enseigné.



A ce stade le constat est clair : les praxéologies vivant dans le savoir savant diffèrent de celles mises en œuvre et enseignées.

Mais pourquoi ? Comment peut-on l'expliquer ? En quoi le NHST est-il adapté à la recherche et à l'enseignement ?

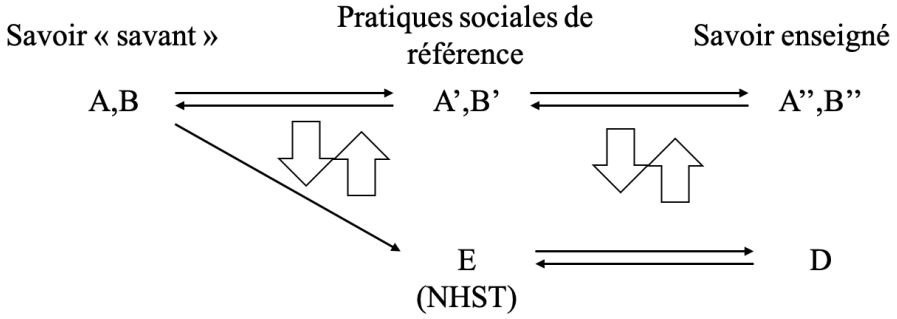
En faisant l'impasse sur l'intégration des éléments subjectifs du contexte, le NHST devient plus facile à mettre en œuvre et semble fournir une démonstration objective de l'existence d'effets expérimentaux.

Le caractère objectif est clairement valorisé dans la recherche dans laquelle la subjectivité est généralement accueillie avec méfiance. Le fait qu'à partir des mêmes données différents chercheurs aboutissent aux mêmes conclusions concernant la mise en évidence ou non d'un effet participe certainement au succès du NHST en recherche. Cette objectivité est également une raison du succès du NHST dans l'enseignement : le NHST est plus facilement *mis en texte*.

Il y a donc, dans l'enseignement comme dans la recherche, des facteurs [2] qui tendent à transformer le test de significativité de Fisher (A) et le test d'hypothèses de Neyman et Pearson (B) en NHST (E).

Si cette tendance est bien présente et a effectivement contribué actuellement à la diffusion du NHST dans les pratiques professionnelles tant que dans l'enseignement, il existe également des facteurs qui tendent à contrer cette tendance. Ceux-ci favoriseraient donc des praxéologies plus en phase avec le savoir savant que l'on pourrait noter A' et B' dans les pratiques sociales de référence et A" et B" dans le savoir enseigné.

C'est le cas de tout ce qui amène un chercheur à s'intéresser au savoir savant : les formations en statistique, les articles remettant en question les pratiques d'inférence statistique, la perception du caractère insatisfaisant des conclusions de recherches basées exclusivement sur le NHST, *etc.* C'est également le cas, en recherche, des *guidelines* telles que les *CONSORT guidelines*. Ces lignes directrices servent de référence pour les chercheurs souhaitant publier le résultat d'essais cliniques, notamment dans les revues biomédicales les plus prestigieuses. Elles contribuent à définir de meilleurs standards en matière de méthodologie scientifique et d'analyse statistique.



En ce qui concerne les outils d'inférence statistique à privilégier dans la présentation des résultats d'essais cliniques on peut y lire :

> "*For all outcomes, authors should provide a confidence interval to indicate the precision (uncertainty) of the estimate. A 95% confidence interval is conventional, but occasionally other levels are used. Many journals require or strongly encourage the use of confidence intervals. They are especially valuable in relation to differences that do not meet conventional statistical significance, for which they often indicate that the result does not rule out an important clinical difference. The use of confidence intervals has increased markedly in recent years, although not in all medical specialties. Although P values may be provided in addition to confidence intervals, results should not be reported solely as P values.*" [Moher et al., 2010]

2. Représentés par une flèche large et vide.

Cette analyse en termes de transposition didactique nous permet de proposer les pistes suivantes pour améliorer l'enseignement de l'inférence statistique :

1. **Changer le savoir enseigné** pour s'approcher de praxéologies plus proches du savoir savant. A ce titre, le test de significativité de Fisher, l'intervalle de confiance ou bien les outils d'inférence bayésiens pourraient constituer des candidats pertinents.
2. Pour se défaire du NHST, il sera nécessaire d'ancrer l'enseignement de l'inférence statistique à la démarche scientifique, de l'**appliquer à de véritables contextes expérimentaux**. En effet, sans partir d'un réel contexte expérimental il semble impossible d'enseigner autre chose que le NHST.
3. Pour soutenir ce changement majeur et le rendre viable dans un contexte où le NHST demeure courant dans les pratiques sociales de référence, il semble utile de **prendre appui sur les pratiques idéales** telles que définies dans des standards qui tendent à s'imposer en recherche biomédicale. Cela devrait, en effet, permettre de conserver un lien fort entre les pratiques professionnelles et le savoir enseigné. Cela impliquerait de centrer l'enseignement de l'inférence statistique autour d'outils d'estimations tel que l'intervalle de confiance plutôt qu'autour du test d'hypothèses.

Dans le chapitre 3, nous interrogeons non plus le *quoi* mais le *comment* : comment enseigner l'inférence statistique ? En effet, au-delà des changements qu'il convient d'apporter au savoir enseigné, des modifications doivent également être apportées à la manière d'aborder celui-ci.

L'analyse *a posteriori* de notre ingénierie didactique se révèle instructive à cet égard car elle permet d'identifier cinq pistes pour améliorer notre enseignement.

— **Un contexte authentique**.
Si l'étude de la transposition didactique nous conduit à souligner l'importance de l'application des outils d'inférence statistique au sein d'un contexte expérimental, l'analyse *a posteriori* de notre ingénierie didactique suggère qu'un pseudo-contexte n'est pas de nature à soutenir l'apprentissage de l'inférence statistique. Pour qu'il puisse jouer ce rôle, le contexte doit être, ou au moins paraitre, authentique.

— **Une consigne accessible**.
Dans le cadre d'une mise en application d'une démarche d'inférence statistique, la tentation est grande d'énoncer la consigne initiale dans les termes des notions visées par l'apprentissage. C'est notamment ce qui a été fait dans notre dispositif expérimental lorsque nous demandions aux étudiants : "*Quelle taille d'échantillon garantit des risques d'erreurs acceptables ?*". Pourtant, la question de la relation entre la taille d'échantillon et le risque d'erreur ne devient accessible pour les étudiants qu'après avoir appris les notions visées, c'est-à-dire bien trop tard. A l'inverse, une consigne plus directe telle que "*Que contient cette bouteille ?*" permet mieux aux étudiants de s'emparer du problème

et d'y engager leurs conceptions.

— **Une démarche claire**.
Pour l'étudiant, le cours de biostatistique est vraisemblablement associé au champ des mathématiques. Il est donc normal qu'il pense que ce que l'enseignant attendu de lui c'est, en premier lieu, qu'il applique des formules correctement. L'enseignant, pour sa part, considèrera que ce qui est attendu de l'étudiant dépasse la simple application de formule : il attend parfois que l'étudiant combine l'application de formules avec la mise en œuvre d'une démarche empirique. Il semble donc nécessaire de prendre en compte ce déséquilibre entre les attentes de l'enseignant et de l'étudiant. Cela pourrait passer par la clarification des attentes au moment de présenter un problème aux étudiants.

— **Un contexte où la population cible est réellement inconnue**.
La conception selon laquelle la population cible est globalement connue figure parmi les conceptions vraisemblables chez nos étudiants, cohérentes avec nos observations et de nature à faire obstacle à l'apprentissage de l'inférence statistique. Dans un contexte où la population cible est connue, il y a une certaine logique à voir la variabilité de l'échantillon comme une erreur et à interpréter les outils d'inférence statistique comme des outils mesurant le niveau auquel l'échantillon choisi est bien représentatif de la population cible. Travailler dans un contexte où les caractéristiques de la population cible sont réellement inconnues semble dès lors nécessaire pour faire évoluer cette conception.

— **Des contextes variés**.
Une deuxième conception pouvant potentiellement constituer un obstacle dans l'apprentissage de l'inférence statistique est la conception selon laquelle la variabilité et les risques d'erreurs ne poseraient problèmes qu'en dessous d'un certain nombre d'observations et que ce nombre serait constant d'une expérience à l'autre. Faire varier les contextes en termes de variabilité observée et de précision recherchée semble dès lors nécessaire pour mettre à mal cette conception.

4.2 Limites

Pour pouvoir aborder le problème de l'enseignement de l'inférence statistique, nous avons opéré une série de choix.

Dans cette thèse, nous avons utilisé des méthodes principalement qualitatives - dans l'analyse épistémologique, lors de l'observation d'étudiants, dans l'analyse de raisonnements d'étudiants, dans la recherche de contraintes institutionnelles, *etc.* - ce qui peut paraître paradoxal lorsque le thème général est l'inférence statistique.

Nous l'expliquons par l'objectif du travail qui, en effet, n'a pas vocation à démontrer la supériorité d'une quelconque approche par rapport à une autre mais cherche plutôt, en amont,

à éclairer les déterminants d'un problème complexe. De cette ébauche de modèle explicatif peuvent naître des hypothèses sur les méthodes qu'il conviendrait de mettre en place pour améliorer l'enseignement de l'inférence statistique à destination des futurs scientifiques et des prédictions sur les savoirs susceptibles d'exister et de survivre au sein de différentes institutions.

L'absence d'analyse quantitative est donc liée au stade assez précoce de la réflexion.

D'autres choix ont probablement occulté une partie du problème mais nous ont permis de mener notre raisonnement à son terme. Il est temps pour nous d'y revenir et d'aborder les questions concernant :

— le fait de travailler sur nos propres étudiants et d'analyser notre propre enseignement ;
— le caractère généralisable d'une analyse qui repose beaucoup sur le niveau local ;
— la manière dont nous avons délimité le champ de l'étude.

4.2.1 A propos de la confusion potentielle entre les postures d'enseignant et de chercheur

En analysant notre propre enseignement et en expérimentant sur nos propres étudiants, nous prenons le risque que la posture de l'enseignant et celle du chercheur interfèrent.

Dans l'étude de la transposition didactique, on peut se demander si l'analyse de notre propre savoir enseigné n'est pas partiale ou biaisée. Nous ne pouvons écarter totalement ce risque même si nous avons tenté de séparer la posture de l'enseignant de celle du chercheur. En effet, il faut noter, d'une part, que ces deux rôles étaient tenus par deux personnes différentes, bien que liées par des liens hiérarchiques (promoteur - doctorant). D'autre part, l'objectif de l'analyse du savoir enseigné a été de le décrire en terme de praxéologie (D) mais pas d'évaluer le travail effectué ou de poser un jugement de valeur. Nous tentons simplement de décrire une certaine transposition didactique et d'identifier les contraintes qui l'influencent.

Par ailleurs, dans le chapitre sur l'ingénierie didactique, nous enregistrons le raisonnement de groupes d'étudiants (par deux ou trois) sélectionnés au hasard dans des groupes de travaux pratiques d'une vingtaine d'étudiants. Dans ce genre d'expérience, si les étudiants associent l'investigateur à l'enseignant (ou à l'assistant), les comportements des étudiants risquent d'être fortement influencés. Nous avons, dès lors, séparé les rôles des différents intervenants : l'enseignant et les assistants n'apparaissaient pas liés à l'investigateur durant l'expérimentation. Par contre, l'expérimentation s'est tout de même déroulée durant les séances de travaux pratiques du cours de statistique ce qui est de nature à influencer, même en l'absence de l'enseignant, la manière dont les étudiants raisonnent.

4.2.2 A propos du caractère généralisable des analyses

Une des particularités de ce travail est d'articuler entre eux deux niveaux : le *local* (cours de statistique à l'Université de Namur, étudiants en sciences biomédicales, pharmacie et médecine, année académique 2018-2019) et un niveau *global* (la recherche scientifique dans les domaines biomédicaux). Il nous semblait, en effet, important de réunir ces deux extrémités afin d'intégrer les besoins du futur milieu professionnel [3] dans la réflexion sur la formation des étudiants.

Cependant, cette restriction dans la définition du niveau *local*, restriction dans l'espace et dans le temps, soulève une série de questions : qu'en est-il dans les autres cours de statistique, donnés ailleurs, à d'autres sections d'étudiants, par des enseignants d'autres formations, dans d'autres contextes ? En quoi ce que nous décrivons dans cette thèse est-il généralisable ?

A cette question, nous pouvons apporter deux éléments de réponse.

D'une part, en tant qu'acteurs d'un système d'enseignement, nous avons une idée générale de la manière dont l'enseignement de l'inférence statistique se déroule classiquement. Cette connaissance est incomplète et informelle mais elle nous permet de penser que la situation n'est pas, ailleurs, déterminée par des contraintes radicalement différentes.

D'autre part, les contraintes institutionnelles que nous avons tenté d'identifier lors de l'analyse de la transposition didactique ont toutes un caractère généralisable, dépassant le niveau purement local. Cela nous laisse penser que ces contraintes peuvent avoir une certaine valeur explicative ou prédictive en dehors du contexte local dont elles sont issues.

4.2.3 A propos de la délimitation du champ d'étude

Une des particularités de notre démarche est de décrire les pratiques professionnelles pour mettre en perspective les difficultés d'enseignement locales. Nous avons tenté de décrire, à la fois, la transposition *didactique* (savoir savant - savoir enseigné) et la transposition *institutionnelle* (savoir savant - pratiques sociales de référence). Ce choix nous semble *a posteriori* tout à fait pertinent mais il a un coût. L'intégration de trois champs différents (recherche scientifique, analyse didactique, notions de statistique) au sein d'une même analyse est, pour nous, une richesse, voire une nécessité, mais a impliqué la réduction du champ d'étude au sein de chaque domaine.

En ce qui concerne la recherche scientifique, par exemple, nous avons présenté une vision très réductrice de la pratique de l'inférence statistique. Nous nous sommes focalisé sur *une* pratique d'analyse des données (le NHST) mais nous n'avons pas pu décrire la diversité des

3. Pour lequel, une fois encore, nous avons forcé le trait puisque tous les étudiants sortant de ces filières ne consacrent pas leur carrière à la recherche scientifique.

approches. Or, la recherche biomédicale est loin de représenter un ensemble homogène, les pratiques d'analyse des données sont très variables notamment selon le type de recherche en jeu. En voici quelques illustrations :

— les recherches en génomique impliquent l'analyse d'une masse considérable de données (des dizaines de milliers de gènes par individu), et ce, dans différentes conditions. Dans ce type d'expérience, le risque de résultat faussement positif est un élément particulièrement important et les méthodes d'inférence statistique sont adaptées en conséquence : seules des P-valeurs extrêmement faibles sont considérées comme rejetant réellement une hypothèse ($P < 10^{-6}$ par exemple) et des notions telles que la *False Discovery Rate* prennent sens pour contrôler le risque d'erreur α ;
— les essais cliniques de phase I [4] impliquent une prise de décision en continu (peut-on augmenter la dose ?) sur base d'un petit nombre de données (quelques volontaires par dose). Dans ce type d'expérience, les méthodes d'inférence bayésienne sont particulièrement adaptées ;
— les méta-analyses combinent, de façon quantitative, les résultats expérimentaux disponibles concernant une question précise. Dans ces analyses, la qualité méthodologique de chaque expérience est prise en compte et le point central de l'analyse est souvent la détermination de l'ampleur d'un effet, via l'intervalle de confiance par exemple, plutôt que la détermination de l'existence ou non d'un effet ;

Les questions qui se posent aux différentes phases de la recherche biomédicale sont variées et les méthodes statistiques mises en œuvre sont, dès lors, variées également. Le fait que nous nous soyons focalisé sur une seule pratique d'inférence statistique dans la recherche constitue, dès lors, une limite de notre travail.

En ce qui concerne l'analyse didactique, nous sommes parti de deux cadres conceptuels qui nous paraissaient *a priori* adaptés, la théorie des situations didactiques de Brousseau et la théorie anthropologique du didactique de Chevallard. Nous avons ainsi fait l'impasse sur la littérature didactique non-francophone et, dans le domaine même de la littérature francophone, sur certaines théories qui, peut-être, auraient été pertinentes pour analyser le problème auquel nous étions confrontés.

Au sein même des cadres conceptuels que nous avons utilisés, nous avons opéré une sélection des concepts en nous focalisant sur ceux qui nous semblaient éclairer le mieux le schéma général. Ainsi avons-nous utilisé les concepts de transposition didactique, de praxéologie, d'obstacle mais n'avons pas parlé de topogenèse, chronogenèse, schéma herbatien, *etc.*

Enfin, en ce qui concerne la statistique, nous avons également restreint le champ d'analyse. Partant du problème de l'enseignement du test d'hypothèses, nous avons tenté de décrire une

4. Dans lesquels des volontaires, sains (en général), reçoivent des doses croissantes d'un médicament en développement afin de déterminer la dose à partir de laquelle des effets secondaires surviennent.

praxéologie associée au test de significativité selon Fisher (A), une praxéologie associée au test d'hypothèses selon Neyman et Pearson (B) et une praxéologie associée au test statistique bayésien basé sur la probabilité *a posteriori* (C). Ces choix sont discutables.

Tout d'abord, décrire le test de significativité selon Fisher par une praxéologie (A) définie par *une* tâche, *une* technique, *une* technologie et *une* théorie relève nécessairement de l'approximation. La vision qu'aurait Fisher du test de significativité dépasse de loin ce que nous avons pu en dire dans la praxéologie A : sa vision est complexe et a pu évoluer au cours du temps ou en fonction de contextes bien précis. Il en va de même pour les autres praxéologies. Il s'agit d'approximations qui, certes, nous semblent justifiées et nécessaires, mais restent des visions assez simplifiées d'une réalité plus complexe.

De plus, si pour la description des outils inférentiels fréquentistes en termes de praxéologie, nous sommes parti de sources directes, cela n'a pas été le cas pour les outils bayésiens pour lesquels nous sommes parti de sources indirectes, finalement plus proches des pratiques sociales de référence que du savoir savant à proprement parler. Bien que cela ne soit pas, de notre point de vue, de nature à modifier substantiellement la nature des conclusions de notre analyse, il aurait été plus cohérent de partir d'une analyse de sources plus proches du savoir savant dans la description des outils bayésiens, par exemple de l'ouvrage *The theory of probability* de Jeffrey (1961) .

Notons également que, dans l'introduction, la définition que nous donnons de l'inférence statistique est probablement assez adaptée à la présentation des tests statistiques paramétriques mais elle occulte, ce faisant, l'ensemble des tests statistiques non-paramétriques.

Enfin, au moment de définir le champ d'analyse, nous avons restreint notre propos aux outils de *test* statistique, excluant ainsi les outils d'*estimation*. Or l'analyse des pratiques en recherche biomédicale révèle que l'intervalle de confiance occupe une place importante en recherche et la lecture des différentes *guidelines* publiées dans cette littérature indique que ces outils sont amenés à occuper une place de plus en plus importante à l'avenir, au détriment des outils de test. Il aurait été intéressant, à ce titre, d'élargir le champ de l'analyse et d'intégrer, une ou plusieurs praxéologies relatives à l'intervalle de confiance.

4.3 Apports et perspectives

Le présent travail trouve ses origines dans un problème d'enseignement et son développement associe un cadre conceptuel et des méthodes issues de la didactique avec une étude critique des pratiques professionnelles. Il nous semble donc opportun de tenter de déterminer les apports de ce travail sur trois plans différents : celui de la biostatistique, celui de la didactique et celui de l'enseignement. Les apports à la biostatistique, à la didactique ou à l'enseignement de manière

générale, seront vraisemblablement difficiles à identifier. Nous espérons que notre réflexion constituera une pierre apportée à l'édifice de la recherche en didactique et qu'elle contribuera à améliorer certaines pratiques mais il est encore trop tôt pour le savoir. En revanche, sur le plan personnel et à une échelle locale, les apports sont tangibles.

4.3.1 Sur le plan de la biostatistique

Sur le plan des pratiques professionnelles, l'analyse épistémologique présentée dans le chapitre 2 se révèle particulièrement éclairante.

En effet, bien que la pratique du NHST soit régulièrement dénoncée, la distinction entre cette pratique et le test de significativité de Fisher ou le test d'hypothèses de Neyman et Pearson ne va pas de soi. Le NHST est souvent présenté comme un hybride entre ces deux outils de test statistique, ce qui entretient une certaine confusion à propos de ce qui distingue réellement ces trois démarches. Notre analyse épistémologique suggère, quant à elle, que la spécificité du NHST repose en réalité sur la mise à l'écart de la démarche scientifique.

Améliorer nos pratiques nécessite donc que l'on replace l'inférence statistique au sein de la démarche scientifique. Cela requiert de prêter plus d'attention aux questions telles que : Quelle est l'hypothèse de recherche sous-jacente à cette expérience ? Est-il possible de la traduire en une hypothèse statistique précise ? Quels seraient les résultats attendus sous cette hypothèse ? Quels écarts peuvent être interprétés comme étant particulièrement important au regard du contexte ? Quelles sont les valeurs de références permettant d'interpréter les résultats *a posteriori*[5], *etc.*

Par ailleurs, l'analyse en termes de contraintes institutionnelles nous éclaire également sur les raisons du succès du NHST dans la recherche biomédicale et nous invite à envisager certaines pistes d'améliorations plutôt que d'autres.

Par exemple, le remplacement pur et simple d'un outil d'inférence statistique – le NHST – par un autre – un test statistique bayésien – ne nous semble pas de nature à résoudre le problème de l'utilisation abusive des outils d'inférence statistique. Par contre, il semble plausible que la diffusion de standards méthodologiques et statistiques parvienne progressivement à renverser le rapport de force et à favoriser la diffusion de pratiques d'inférence statistique moins dommageables pour la science.

Chercher à mieux enseigner l'inférence statistique nous amène donc, indirectement, à mieux comprendre les outils que nous manipulons quotidiennement en analysant des données dans le cadre de recherches biomédicales et, de ce fait, à faire évoluer notre pratique.

5. Et éviter ainsi la seule interprétation possible en l'absence de tels points de repère : "l'effet est statistiquement significatif" ou "l'effet n'est pas statistiquement significatif".

4.3.2 Sur le plan de la didactique

Montrer les apports de cette thèse sur le plan de la didactique nécessite de faire référence à notre point de départ – l'état de notre réflexion au début de la thèse en l'occurrence – et de montrer la progression qui a été la nôtre tout au long de ce travail.

Partant d'un travail de recherche antérieur[Calmant, 2004], nous avons, tout au long de notre réflexion modifié considérablement notre point de vue, et ce, à divers égards.

Dès nos premières expérience d'enseignement de la biostatistique lors des séances de travaux pratiques, nous avons constaté les difficultés qu'avaient les étudiants à "maitriser la matière". Nous avons alors interprété la présence de ces difficultés par un manque de motivation de la part de ces étudiants. Peut-être ont-ils du mal à apprendre car ils ne voient pas à quoi servent les notions enseignées ? Nous avons alors entrepris d'étudier la dynamique motivationnelle de ces étudiants concernant le cours de biostatitique [Bihin, 2013].

Mais, les difficultés des étudiants semblent, en fait, se cristalliser autour de l'application et de l'interprétation de tests statistiques, lorsque l'on aborde l'inférence statistique. Nous nous sommes donc intéressés aux difficultés liées à l'apprentissage du test d'hypothèses.

Nous avons alors réalisé des enquêtes sur le thème de l'inférence statistique (et, plus précisément de la P-valeur) auprès de ces étudiants afin d'objectiver les difficultés que nous pressentions. A ce stade, nous interprétions leurs erreurs comme l'absence d'une conception, un simple manque qu'il conviendrait de combler en énonçant de manière plus claire les notions problématiques.

Nous avons ensuite réalisé, en commençant à travailler avec des médecins dans le domaine de la recherche biomédicale, que les 'erreurs' que nous décrivions chez les étudiants se retrouvent aussi largement chez les chercheurs. Un examen rapide de la littérature à ce sujet nous apprend que ces erreurs sont régulièrement dénoncées. Le problème de la 'maitrise de l'inférence statistique' ne semble donc pas spécifique d'une cohorte d'étudiants, d'un cours ou d'un groupe de chercheurs mais semble résider dans le savoir en lui-même ou dans la manière dont la théorie est appliquée.

Cela nous a amené à poser la question de l'origine du savoir enseigné. Que disent les théories initiales que nous prétendons enseigner ? S'en distingue-t-on ? L'étude de la transposition didactique nous révèle la nature des écarts existant entre le savoir savant et le savoir enseigné.

Émerge alors une nouvelle question : l'écart au savoir savant serait-il la cause des difficultés d'apprentissage chez les étudiants ? Les étudiants auraient-ils plus facile à apprendre si l'on enseignait un savoir plus fidèle au savoir savant ?

Nous avons voulu tester cette hypothèse à l'aide de l'ingénierie didactique. Brousseau nous

invite, en effet, à voir les erreurs non comme un manque mais comme la marque de la présence d'une conception qui a, nécessairement, déjà fait ses preuves. Pour favoriser l'apprentissage, il convient alors de placer les étudiants face à des situations dans lesquelles la conception visée est censée être une réponse bien plus satisfaisante, plus économique que les conceptions déjà existantes. Pour cela, nous avons conçu un dispositif expérimental censé permettre l'apprentissage de la logique du test d'hypothèses, dans une version plus proche du savoir savant. La mise en œuvre et l'analyse *a posteriori* ne purent corroborer notre hypothèse, et suggérèrent que le simple remplacement du savoir enseigné par le savoir savant ne permettrait pas de résoudre les difficultés d'apprentissage qui nous occupent.

En fait, peut-être que décrire les écarts entre le savoir enseigné et le savoir savant en termes d'erreur n'est pas judicieux. Chevallard nous propose, en effet, d'analyser le problème en d'autres termes. Il conviendrait, selon lui, de voir le savoir enseigné comme une réponse adaptée aux contraintes qui agissent au sein d'une certaine institution, contraintes qu'il est nécessaire d'identifier si l'on souhaite imaginer d'autres réponses. La question devient alors : quelles sont ces contraintes auquel le savoir est soumis ?

Indéniablement, le savoir enseigné dans un cours de biostatistique appliquée se doit d'être suffisamment proche des pratiques professionnelles, la recherche biomédicale en l'occurrence.

Se pose alors la question des pratiques sociales de référence : quelles sont-elles ? Elles sont certainement diverses et rarement explicitée dans les articles scientifiques. Toutefois, dans la littérature biomédicale on retrouve des auteurs qui dénoncent les écarts entre les théories initiales (de Fisher, Neyman et Pearson) et une pratique courante, le NHST.

Cohen (1994) montre que ces pratiques tentent de remplacer la logique de corroboration d'une hypothèse par une logique de démonstration probabiliste par l'absurde qui n'est pas valide. Carver (1978) dénonce le fait que ces pratiques corrompent la démarche scientifique : déterminer la significativité statistique d'un effet élude trop souvent, selon lui, la définition d'une hypothèse de recherche précise, l'identification des résultats prédits par cette hypothèse et l'interprétation d'un effet dans son contexte. Gigerenzer (2004) parle, lui, de ces pratiques comme d'un rituel absurde et met en garde contre la recherche d'une méthode d'inférence statistique universelle.

On pourrait s'arrêter à ces constats et interpréter ces écarts comme des erreurs. Mais il semble, une nouvelle fois, plus intéressant de voir ces pratiques comme des réponses adaptées à un certain environnement. Cela nous amène à chercher les contraintes institutionnelles qui agissent sur les pratiques professionnelles.

Dans le contexte qui caractérise la recherche scientifique, le NHST semble particulièrement adapté. Il semble constituer une méthode "tout-terrain", objective, produisant des messages clairs et éludant certaines difficultés liées à la démarche scientifique. Le NHST semble également

bien adapté à l'enseignement car il est facilement *mis en texte* comme dirait Chevallard.

Cette analyse permet donc de comprendre les déterminants à l'œuvre et, ce faisant, d'éclairer une question importante pour l'enseignement et pour la recherche scientifique.

D'un simple constat de "matière non maitrisée" et d'étudiants "peu motivés" lors des travaux pratiques nous arrivons, au fil de notre raisonnement, à proposer une hypothèse originale et vraisemblable quant aux facteurs qui agissent sur les pratiques professionnelles et, par conséquent, sur le savoir enseigné, et qui expliquent la tendance qu'on ceux-ci à dévier des théories initiales.

L'hypothèse que nous formulons concernant les contraintes institutionnelles favorisant la diffusion du NHST demande à être consolidée (au vu notamment des limites identifiées) et éprouvée. Pour ce faire, il sera nécessaire d'en déduire des conséquences logiques et de les confronter aux observations afin de voir si ces dernières corroborent cette hypothèse.

Par exemple, si cette hypothèse est correcte, on peut s'attendre à ce que les éléments subjectifs du test de significativité (interprétation de la P-valeur en fonction du contexte) ou du test d'hypothèses (définition des risques acceptables en fonction du contexte) ne se retrouvent pas dans le savoir enseigné dans d'autres établissements similaires au nôtre. On peut également prédire que, dans ces établissements, le savoir enseigné ne sera pas constitué d'une véritable boîte à outils d'inférence statistique mélangeant les approches d'estimation et de test, fréquentistes et bayésiennes. Dans le même ordre d'idées, si les méthodes bayésiennes étaient enseignées, alors le caractère subjectif du choix des (distributions de) probabilité *a prior* serait compliqué à transmettre au-delà de simples règles de conduites applicables en tout contexte. Enfin, on peut prédire que si l'enseignant tente de réorganiser l'enseignement de l'inférence statistique autour d'un outil d'estimation tel que l'intervalle de confiance plutôt que des outils de test, alors il sera confronté à la difficulté d'enseigner comment interpréter ces intervalles en fonction du contexte, au-delà du simple constat de la cohérence ou non des observations avec une hypothèse nulle ce qui reviendrait *de facto* à appliquer la démarche sous-jacente au NHST.

Le caractère général de notre hypothèse et des prédictions qui peuvent en découler en fait un matériau intéressant pour la construction de connaissances en didactique de la statistique.

4.3.3 Sur le plan de l'enseignement

Enfin, nous pouvons noter que la présente recherche a déjà profondément transformé nos pratiques sur le plan de l'enseignement. Elle suggère, en effet, une modification importante du savoir enseigné et de la manière de l'aborder.

Durant l'année académique 2020-2021, nous avons eu l'opportunité de mettre en œuvre ce changement à l'occasion d'une réforme du cours de biostatistique donné aux étudiants de

première année de médecine. Pour ces étudiants, en effet, le cours de biostatistique est passé d'une unité d'enseignement (MBIOB131, 4 ECTS) comportant 26 h théoriques et 15 h de travaux pratiques à deux unités d'enseignement : l'une comprenant 26 h théoriques (MBIO1B32, 2 ECTS) et l'autre comprenant 26 h théoriques + 16 h de travaux pratiques (MBIOB133, 2 ECTS).

Nous avons tenté de changer radicalement la matière enseignée ainsi que l'approche adoptée (voir tableau 4.1).

Cette réforme repose sur deux changements majeurs.

Le premier élément est la division du cours en deux phases caractérisées par des attentes différentes concernant le type de démarche que l'étudiant doit mettre en œuvre. Durant la unité d'enseignement (MBIOB132 donné au premier quadrimestre), l'enseignement est assez classique dans la mesure où il implique une présentation *ex cathedra* des notions, l'illustration de ces notions avec des exemples issus de la littérature biomédicale, l'utilisation d'exercices de calculs et une évaluation écrite à l'aide de questions fermées.

Ensuite, dans la deuxième unité d'enseignement (MBIOB133 donné au deuxième quadrimestre), l'enseignement s'organise autour de la mise en pratique des notions de statistique descriptive et inférentielle au sein d'une démarche expérimentale. En parallèle à la présentation *ex cathedra* et l'illustration de notions de statistique et de méthodologie expérimentale, les étudiants sont engagés dans un travail de groupe s'étendant sur les 16 h de travaux pratiques. Ce travail est l'occasion, pour les étudiants, de mettre en application les notions théoriques dans un contexte de recherche. Les étudiants doivent s'inspirer de la méthodologie des essais cliniques pour élaborer un protocole d'expérience, le mettre en œuvre, récolter et encoder les observations, réaliser les analyses statistiques adéquates et interpréter les résultats. La question de départ est choisie par les étudiants et énoncée en termes simples tel que "*A quel point la vitesse de lecture est-elle plus importante sur papier que sur écran ?*". Ce genre de question de recherche implique donc un contexte authentique dans lequel les caractéristiques de la population cible sont réellement inconnues : les étudiants ignorent véritablement le niveau auquel la vitesse de lecture pourrait être impactée par le support.

Le deuxième élément est le changement dans les outils inférentiels présentés. Là où l'enseignement reposait principalement sur le test d'hypothèses [6] et la P-valeur, il repose désormais principalement sur l'intervalle de confiance et, dans une moindre mesure, sur la P-valeur. Ce changement s'appuie sur les recommandations décrites dans les *guidelines CONSORT*. Il est à noter que le doublement du nombre d'heures théoriques n'a pas été accompagné d'une augmentation dans la quantité de thèmes abordés, au contraire. Cette augmentation du volume horaire a permis de dégager du temps pour la mise en place d'une démarche expérimentale.

6. dans une version différente du test d'hypothèses de Neyman et Pearson comme nous avons pu le voir.

TABLE 4.1 – **Evolution du cours de biostatistique en médecine entre 2017-2018 et 2020-2021**. *V* : notions abordées au cours ("vues"), *C* : notions impliquant des exercices avec calculs, *I* notions illustrées par des exemples pris dans la littérature biomédicale, *A* notions potentiellement appliquées dans une démarche de recherche.

		2017-2018				2020-2021			
		V	C	I	A	V	C	I	A
Méthodologie expérimentale	Design expérimental	■		■		■		■	■
	Randomisation	■	■	■		■	■	■	■
	Conduite en aveugle	■		■		■		■	■
	Plan d'analyse statistique	■		■		■		■	■
	Gestion des données manquantes	■		■		■		■	■
Statistique descriptive	Tendance centrale et dispersion	■	■	■		■	■	■	■
	Histogramme	■	■	■		■	■	■	■
	Méthode de Kaplan-Meier	■	■	■		■	■	■	■
	Méthode actuarielle	■	■	■					
	Coefficient de corrélation linéaire	■	■			■	■	■	■
	Droite des moindres rectangles, linéarisation de relations	■							
	Tests diagnostiques	■	■	■					
Probabilités	Evènement, probabilités conditionnelles	■	■						
Distributions théoriques	Exponentielle	■	■	■		■	■	■	■
	Binomiale	■	■						
	Poisson	■	■						
	Normale	■	■			■	■	■	■
	Student	■	■						
	Chi-Carré	■	■						
	Fisher	■	■						
Test d'hypothèses	Principe	■		■		■		■	
	Calcul de taille d'échantillon	■	■	■					
	Test de Hartley	■	■						
	Test de Fisher	■	■						
	Test de Z	■	■						
	Test de t simple et pairé	■	■						
	ANOVA	■	■						
	Test du log-rank	■	■						
P-valeur	Principe	■	■	■		■	■	■	■
Intervalles de confiance	Principe	■		■		■	■		
	Proportion					■	■	■	■
	Différence de proportion					■	■	■	■
	Moyenne					■	■	■	■
	Différence de moyenne					■	■	■	■
	Odds ratio					■	■	■	■

Les premiers retours de cette réforme sont très encourageants, ils témoignent de la faisabilité d'une telle réforme même dans un contexte où la taille des cohortes d'étudiants (ici environ 300 étudiants) constitue habituellement un obstacle à un enseignement impliquant l'expérimentation. Les étudiants se sont visiblement engagés dans la tâche comme attendu, ils ont véritablement pris le rôle du chercheur préparant son expérience, élaborant son protocole, opérant des choix dans l'analyse statistique et essayant de répondre à une question de recherche à

partir de données expérimentales soumises à une certaine variabilité.

Ce travail appliqué a permis de soulever les questions donnant sens aux notions d'inférence statistique et que les étudiants ne se posent normalement pas avant d'avoir à faire de la recherche, c'est-à-dire bien après les cours d'introduction à l'inférence statistique. Parmi ces questions, nous trouvons notamment : "*Combien d'individu faut-il inclure dans cette expérience pour que celle-c fournisse une estimation suffisamment précise de l'effet étudié ?*" ou bien "*A quel point ces observations sont-elles compatibles avec l'hypothèse selon laquelle l'effet de l'intervention serait important ?*".

La mise en œuvre de cette réforme soulève de nouvelles questions pour l'enseignement de l'inférence statistique.

On peut, tout d'abord, se demander comment améliorer le dispositif d'enseignement actuel. La sélection des notions enseignée est-elle pertinente ? Quel niveau de contrainte faut-il imposer à la démarche expérimentale imposée aux étudiants ? Comment évaluer une telle démarche ? Comment ce dispositif d'enseignement va-t-il vieillir ?

Il sera, ensuite, intéressant de déterminer dans quelle mesure ce dispositif d'enseignement améliore la maitrise que ces étudiants peuvent avoir de l'inférence statistique. Quelles sont les conceptions faisant obstacle à l'apprentissage de l'inférence statistique dans un tel contexte ?

Enfin, il faudra également s'interroger sur le caractère généralisable des pistes qui sont proposées dans cette thèse et mises en œuvre dans cette réforme.

Un cours de statistique appliquée doit nécessairement s'adapter aux spécificités de chaque filière et aux contingences locales (nombre d'étudiants, bagage mathématique ou informatique, objectifs de la formation, *etc.*). A travers la réforme présentée ici nous n'avons donc pas la prétention de proposer une solution universelle mais, plutôt, de présenter une piste pour tenter d'améliorer l'enseignement de l'inférence statistique chez les étudiants des filières biomédicale.

Celle-ci implique, d'une part, de remplacer le test d'hypothèses par l'intervalle de confiance dans le savoir enseigné et, d'autre part, de dégager de la place, au sein du cours de statistique, pour mettre en œuvre une véritable démarche de recherche avec les étudiants, soit sous la forme d'expériences proches des essais cliniques, soit par la pratique d'enquêtes, ou par tout ce qui peut contribuer à appliquer les méthodes statistiques dans le cadre d'un problème concret. Il importe d'essayer, le plus souvent possible, de ne pas séparer la statistique en général, et l'inférence statistique en particulier, de son substrat expérimental.

En cela, nous rejoignons Régnier sur la nécessisté d'une *mise en œuvre effective des concepts et techniques au travers d'une situation problème de statistique appliquée* [Régnier, 2005]. Ceci rejoint également l'appel de Batanero et Díaz à enseigner la statistique à partir de problèmes concrets et en lien avec la recherche scientifique :

Statistical inference is just a part of the more general process of scientific inference. However, we frequently teach statistics in isolation without connecting it with a more general framework of research methodology and experimental design. From our point of view, it is necessary to discuss the role of statistics within experimental research with the students and make them conscious of the possibilities and limitations of statistics in experimental work [Batanero and Díaz, 2008].

Epilogue

Qui est coupable dans l'affaire dite des difficultés d'enseignement de l'inférence statistique ?

Dans notre enquête, la motivation des étudiants envers le cours de biostatistique est la première suspecte mais les enquêtes préliminaires permettent rapidement de l'innocenter. Le problème n'est pas, en effet, propre à l'ensemble du cours de biostatistique mais plutôt à une partie bien spécifique de celui-ci : l'inférence statistique.

Vient le tour du savoir enseigné localement. Est-il vraiment fidèle au savoir savant ? Les écarts par rapport au savoir savant expliquent-ils les difficultés des étudiants ? Certes, les écarts s'avèrent importants mais le savoir enseigné bénéficie de plusieurs circonstances atténuantes : (1) l'infidélité au savoir savant est, apparemment, un problème récurrent avec les savoirs enseignés, (2) rien ne permet de penser que la fidélité au savoir savant permettrait de réduire les difficultés des étudiants, et (3) tout indique que le savoir enseigné a agit sous l'influence des pratiques sociales de référence.

Complice donc, mais pas seul responsable.

Mais quelles sont ces fameuses pratiques qui, pourtant inlassablement dénoncées, continuent à sévir dans le milieu de la recherche scientifique et à imposer leur modèle à l'enseignement ? D'autres enquêteurs, déjà sur le coup, nous mettent sur la piste de leur suspect : un outil dangereux, du nom de NHST, particulièrement dommageable pour la recherche scientifique.

L'analyse biométrico-praxéologique est formelle : il s'agit du même outil que celui incriminé dans l'enseignement. Les preuves concordent et le suspect est emmené au tribunal.

On l'accuse de corrompre, à grande échelle, la démarche scientifique. Il empêcherait les chercheurs de raisonner correctement et les étudiants de comprendre l'inférence statistique qu'il est, pourtant, censé servir !

"Traitre ! Vendu !", entend-on dans la salle.

Le juge s'apprête à déclarer, une nouvelle fois, la culpabilité du NHST.

Mais un élément l'intrigue...

Attendez.

Ne vient-on pas de dire que le NHST est un outil ?

Et si ce n'était que l'arme du crime ?

Dans ce cas, où sont les coupables ?

D'ailleurs, que font ces chercheurs qui, sans le moindre alibi scientifique, déambulent sur la scène du crime ?

N'ont-ils pas, eux-aussi, leur part de responsabilité ? Qu'est-ce qui les amène à déléguer de si lourdes décisions à un simple outil statistique ?

Comment la communauté scientifique a-t-elle pu tolérer, et parfois encourager, des pratiques dénaturant à ce point la démarche scientifique ?

Le juge commence à douter et le temps vient à manquer.

L'audience est suspendue mais l'enquête, elle, suit son cours...

Bibliographie

[Altman and Bland, 1994a] Altman, D. G. and Bland, J. M. (1994a). Statistics Notes : Diagnostic tests 1 : sensitivity and specificity. *BMJ*, 308(6943) :1552–1552.

[Altman and Bland, 1994b] Altman, D. G. and Bland, J. M. (1994b). Statistics Notes : Diagnostic tests 2 : predictive values. *BMJ*, 309(6947) :102–102.

[Armitage and Colton, 1998] Armitage, P. and Colton, T., editors (1998). *Encyclopedia of biostatistics*. J. Wiley, Chichester ; New York.

[Artigue, 1988] Artigue, M. (1988). Ingénierie didactique. *Recherches en didactique des mathématiques*, 9(3) :281–308.

[Artigue, 1989] Artigue, M. (1989). Ingénierie didactique. *Publications mathématiques et informatique de Rennes*, (S6) :124–128.

[Artigue, 1999] Artigue, M. (1999). The teaching and learning of mathematics at the university level : Crucial questions for contemporary research in education. *Notices of the American Mathematics Society*, 46 :1377–1385.

[Artigue, 2002] Artigue, M. (2002). Ingénierie didactique : quel rôle dans la recherche didactique aujourd'hui ? *Les dossiers des sciences de l'éducation*, 8(1) :59–72.

[Astolfi, 1990] Astolfi, J.-P. (1990). Les concepts de la didactique des sciences, des outils pour lire et construire les situations d'apprentissage. *Recherche et Formation*, (8) :19–31.

[Bachelard, 1934] Bachelard, G. (1934). *La formation de l'esprit scientifique : contribution à une psychanalyse de la connaissance objective*. Bibliothèque des textes philosophiques. Vrin, Paris, 5e édition (1967) edition.

[Baggerly and Coombes, 2009] Baggerly, K. A. and Coombes, K. R. (2009). Deriving chemosensitivity from cell lines : Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 3(4) :1309–1334.

[Baker, 2016] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604) :452–454.

[Batanero and Díaz, 2008] Batanero, C. and Díaz, C. (2008). Methodological and Didactical Controversies around Statistical Inference.

[Berger, 2003] Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing ? *Statistical Science*, 18(1) :1–32.

[Berry, 2006] Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1) :27–36.

[Bihin, 2013] Bihin, B. (2013). *Dynamique motivationnelle des étudiants dans l'apprentissage de la biostatistique : analyse exploratoire et conception d'un questionnaire.* Mémoire présenté en vue de l'obtention du Master complémentaire en pédagogie universitaire et de l'enseignement supérieur., Université catholique de Louvain : Faculté de psychologie et des sciences de l'éducation, école d'éducation et de formation., Louvain-la-Neuve.

[Box, 1979] Box, G. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics*, pages 201–236. Elsevier.

[Brousseau, 1986] Brousseau, G. (1986). *Théorisation des phénomènes d'Enseignement des Mathématiques.* Thèse d'état, Université Bordeaux 1, Bordeaux, France.

[Brousseau, 1997] Brousseau, G. (1997). *Theory of didactical situations in mathematics : didactique des mathématiques, 1970-1990*, volume 19 of *Mathematics education library.* Kluwer Academic Publishers, Dordrecht ; Boston.

[Brousseau, 1998] Brousseau, G. (1998). Les obstacles épistémologiques, problèmes et ingénierie didactique. In *La théorie des situations didactiques*, Recherches en Didactiques des Mathématiques., pages 115–160. Grenoble, France, la pensée sauvage edition.

[Brousseau, 2000] Brousseau, G. (2000). Educacion y Didactica de las mathématicas. *Education matematica (Mexico)*, 12(1) :5–39.

[Brousseau, 2005] Brousseau, G. (2005). Une expérience de premier enseignement des statistiques et des probabilités. In Mercier, A. and Margolinas, C., editors, *Balises en didactique des mathématiques : Cours de la 12e École d'été de didactique des mathématiques*, pages 165–249. Grenoble, France, la pensée sauvage edition.

[Brousseau, 2010] Brousseau, G. (2010). Glossaire de quelques concepts de la théorie des situations didactiques en mathématiques (1998).

[Brousseau et al., 1974] Brousseau, G., Briand, J., Brousseau, N., and Llorens, M. (1974). Description des 31 leçons expérimentées à l'école J. Michelet à Talence. *Compte-rendu de la 26e rencontre de la CIEAEM ; Bordeaux août 1974*, pages 82–123.

[Calmant, 2004] Calmant, P. (2004). *Favoriser l'apprentissage des biostatistiques par le Web ? Essai de problématisation didactique d'une question issue du terrain.* Dissertation en vue de l'obtention du grade de Docteur en Sciences, Facultés Universitaires Notre-Dame de la Paix, Namur.

[Calmant et al., 2017] Calmant, P., Vincke, G., Wauthy, A.-C., and Depiereux, E. (2017). Pratique des biostatistiques.

[Carver, 1978] Carver, R. (1978). The Case Against Statistical Significance Testing. *Harvard Educational Review*, 48(3) :378–399.

[Chevallard, 1989] Chevallard, Y. (1989). On Didactic Transposition Theory : Some Introductory Notes. *Proceedings of the International Symposium on Selected Domains of Research and Development in Mathematics Education (Bratislava)*, pages 51–62.

[Chevallard, 1991] Chevallard, Y. (1991). *La transposition didactique : du savoir savant au savoir enseigné*. Recherches en didactique des mathématiques. La Pensée Sauvage, Grenoble, 2. éd edition. OCLC : 257525638.

[Chevallard, 1997] Chevallard, Y. (1997). Les savoirs enseignés et leurs formes scolaires de transmission : un point de vue didactique. *Skholê*, (7) :45–64.

[Chevallard, 1998] Chevallard, Y. (1998). Analyse des pratiques enseignantes et didactique des mathématiques : l'approche anthropologique. *Actes de l'université d'été de l'IREM de Clermont-Ferrand : Analyse des pratiques.*, pages 91–120.

[Chevallard, 2007] Chevallard, Y. (2007). Passé et présent de la théorie anthropologique du didactique. *Actes du congrès international sur la théorie anthropologique du didactique L. Ruiz-Higueras, A. Estepa, & F. Javier García (Éd.), Sociedad, Escuela y Mathemáticas. Aportaciones de la Teoría Antropológica de la Didáctico, Universidad de Jaén.*, pages 705–746.

[Chevallard, 2013] Chevallard, Y. (2013). Éléments de théorie anthropologique du didactique (TAD). Une initiation à la didactique fondamentale.

[Christensen, 2005] Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59(2) :121–126.

[Cohen, 1994] Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12) :997–1003.

[Collins English Dictionary, 2020] Collins English Dictionary (2020). https ://www.collinsdictionary.com/dictionary/english/descriptive-statistics.

[Cox, 1958] Cox, D. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, pages 357–372.

[Dagnelie, 2010] Dagnelie, P. (2010). Dans l'enseignement de la statistique, la simulation a des limites. *Statistique et Enseignement*, 1(2) :67–68.

[De Finetti, 1974] De Finetti, B. (1974). *Theory of Probability*, volume 1. New York, john wiley and sons edition.

[Depiereux, 2016] Depiereux, E. (2016). *De la variabilité aux risques d'erreur : analyse critique des résultats expérimentaux et de la fiabilité d'une décision clinique*. Presses universitaires de Namur, Namur. OCLC : 962730178.

[Dodge, 2003] Dodge, Y., editor (2003). *The Oxford dictionary of statistical terms*. Oxford University Press, Oxford ; New York, 6th ed edition. OCLC : ocm52324277.

[Fisher, 1925] Fisher, R. A. (1925). *Statistical methods for research workers.* Oliver Boyd, Edinburgh, fifth edition.

[Fisher, 1929] Fisher, R. A. (1929). The statistical method in psychical research. *Proceedings of the Society fo Psychical Research*, 39 :189–192.

[Fisher, 1955] Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society : Series B (Methodological)*, 17(1) :69–78.

[Gigerenzer, 2004] Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5) :587–606.

[Gigerenzer and Marewski, 2015] Gigerenzer, G. and Marewski, J. N. (2015). Surrogate Science : The Idol of a Universal Method for Scientific Inference. *Journal of Management*, 41(2) :421–440.

[Goodman, 1999] Goodman, S. N. (1999). Toward Evidence-Based Medical Statistics. 1 : The P Value Fallacy. *Annals of Internal Medicine*, 130(12) :995.

[Herndon et al., 2014] Herndon, T., Ash, M., and Pollin, R. (2014). Does high public debt consistently stifle economic growth ? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2) :257–279.

[Ioannidis, 2005a] Ioannidis, J. P. A. (2005a). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA*, 294(2) :218.

[Ioannidis, 2005b] Ioannidis, J. P. A. (2005b). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8) :e124.

[Jeffreys, 1961] Jeffreys, H. (1961). *Theory of probability.* Oxford University Press, Oxford [Oxfordshire] : New York, 3rd ed edition.

[Kass, 2011] Kass, R. E. (2011). Statistical Inference : The Big Picture. *Statistical Science*, 26(1) :1–9.

[Kilkenny et al., 2010] Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., and Altman, D. G. (2010). Improving Bioscience Research Reporting : The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biology*, 8(6) :e1000412.

[Koliopoulos et al., 2013] Koliopoulos, D., Boilevin, J.-M., Dossis, S., Paraskevopoulou, E., and Ravanis, K. (2013). Rapport au savoir scientifique de futurs professeurs des écoles en France et en Grèce : le cas du pendule. *RDST. Recherches en didactique des sciences et des technologies*, (8) :163–188.

[Kowalski, 2010] Kowalski, C. J. (2010). Pragmatic Problems with Clinical Equipoise. *Perspectives in Biology and Medicine*, 53(2) :161–173.

[Margolinas, 2005] Margolinas, C. (2005). Essai de généalogie en didactique des mathématiques. *Société suisse pour la recherche en éducation (SSRE)*, 27(3) :343–360.

[Martinand, 1989] Martinand, J.-P. (1989). Pratiques de référence, transposition didactique et savoirs professionnels en sciences techniques. *Les sciences de l'éducation, pour l'ère nouvelle*, 2 :23–29.

[Mayo and Spanos, 2006] Mayo, D. G. and Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science*, 57(2) :323–357.

[Moher et al., 2010] Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., and Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration : updated guidelines for reporting parallel group randomised trials. *BMJ*, 340(mar23 1) :c869–c869.

[Moore, 2007] Moore, D. S. (2007). *The basic practice of statistics*. Freeman, New York, 4. ed edition. OCLC : 254369595.

[Nau, 2001] Nau, R. F. (2001). De Finetti was Right : Probability Does Not Exist. *Theory and Decision*, 51(2/4) :89–124.

[Neyman and Pearson, 1928] Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20A(3-4) :263–294.

[Neyman and Pearson, 1933] Neyman, J. and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, 231 :289–337.

[Nuzzo, 2014] Nuzzo, R. (2014). Scientific method : Statistical errors. *Nature*, 506(7487) :150–152.

[Park, 2012] Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential Reasoning in Statistics : An Argument-Based Approach to Validation*. PhD thesis, The university of Minnesota, Minnesota.

[Peng, 2015] Peng, R. (2015). The reproducibility crisis in science : A statistical counterattack. *Significance*, 12(3) :30–32.

[Popper, 1935] Popper, K. (1935). *Logik der Forschung*. Julius Springer, Autriche, Vienne.

[Potti et al., 2011] Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., Cragun, J., Cottrill, H., Kelley, M. J., Petersen, R., Harpole, D., Marks, J., Berchuck, A., Ginsburg, G. S., Febbo, P., Lancaster, J., and Nevins, J. R. (2011). Retraction Note : Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, 17(1) :135–135.

[Romeijn, 2017] Romeijn, J.-W. (2017). Philosophy of Statistics.

[Régnier, 1998] Régnier, J.-C. (1998). La prise de décision risquée en situation incertaine : élément pour une séquence didactique visant l'acquisition du raisonnement statistique. Enseigner la Statistique du CM à la Seconde Pourquoi ? Comment ? Technical report, IREM de Lyon - Université Lyon1. halshs-00406126.

[Régnier, 2005] Régnier, J.-C. (2005). Formation de l'esprit statistique et raisonnement statistique. Que peut-on attendre de la didactique de la statistique ? *Séminaire National de Didactique des Mathématiques*, pages 13–38.

[Régnier, 2012] Régnier, J.-C. (2012). Enseignement et apprentissage de la statistique : entre un art pédagogique et une didactique scientifique. *Statistique et Enseignement*, 3(1) :19–36.

[Thompson, 2001] Thompson, B. (2001). 402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies.

[Vandenbroucke et al., 2007] Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., and STROBE Initiative (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) : explanation and elaboration. *Epidemiology (Cambridge, Mass.)*, 18(6) :805–835.

[Vincke et al., 2014] Vincke, G., Bihin, B., Wauthy, A.-C., Al Zind, E., Vervoort, A., and Depiereux, E. (2014). Suivi des apprentissages au moyen d'évaluations formatives par questions à choix multiples diffusées sur le Web par le logiciel eTests. *Revue internationale des technologies en pédagogie universitaire*, 11(1) :19–34.

[Vincke et al., 2013] Vincke, G., Wauthy, A.-C., Bihin, B., and Depiereux, E. (2013). Quand la délocalisation numérique d'une partie d'un dispositif d'apprentissage permet de recentrer le temps présentiel sur un obstacle : exemple de l'appropriation de la courbe de Gauss par la manipulation d'objets concrets. : Recentrer le temps présentiel sur un obstacle. *Revue internationale des technologies en pédagogie universitaire*, 10(1) :16–28.

[Wasserstein and Lazar, 2016] Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's Statement on p -Values : Context, Process, and Purpose. *The American Statistician*, 70(2) :129–133.

Annexes

TABLE 4.2: Association entre les racines et les concepts

	Racine	Concept
1	ambigu	Ambiguite
2	ambiguit	Ambiguite
3	ambigus	Ambiguite
4	appliqu	Formule
5	ball	Dispositif
6	bill	Dispositif
7	binomial	Binomiale
8	blanc	Dispositif
9	blanch	Dispositif
10	boug	Variabilite
11	boul	Dispositif
12	brochet	Estimation
13	calcul	Formule
14	chapitr	Cours
15	chinois	Difficulte
16	combien	Quantite
17	combin	Combinaison
18	combinaison	Combinaison
19	combinatoir	Combinaison
20	compatibl	Compatibilite
21	compliqu	Difficulte
22	context	Contexte
23	cour	Cours
24	cout	Contexte
25	croix	Dispositif
26	diagnostic	Contexte
27	difficil	Difficulte

28	distribu	Binomiale
29	ecart	Variabilite
30	echantillon	Echantillon
31	echec	Binomiale
32	ennui	Difficulte
33	epidem	Contexte
34	erreur	Erreur
35	essai	Experience
36	experient	Experience
37	experiment	Experience
38	experimental	Experience
39	facil	Difficulte
40	fluctuent	Variabilite
41	formul	Formule
42	formulair	Formule
43	fourchet	Estimation
44	gripp	Contexte
45	grippal	Contexte
46	h1n1	Contexte
47	h2n3	Contexte
48	h3n1	Contexte
49	h3n2	Contexte
50	h3n210	Contexte
51	h3n3	Contexte
52	hypoth	Hypothese
53	hypothes	Hypothese
54	impossibl	Possible
55	imprecis	Precision
56	improb	Probabilite
57	independ	Binomiale
58	individus	Tirage
59	intervall	Intervalle
60	labos	Contexte
61	lanc	Tirage
62	lancer	Tirage
63	machin	Truc
64	malad	Contexte
65	malaid	Contexte
66	malchanc	Probabilite

67	marg	Estimation
68	medecin	Contexte
69	mesur	Quantification
70	nombr	Nombre
71	null	Contexte
72	patient	Contexte
73	perdu	Difficulte
74	perplex	Difficulte
75	perturb	Difficulte
76	pi0	Pi
77	pi02	Pi
78	pi04	Pi
79	pi1	Pi
80	pi15	Pi
81	pi20	Pi
82	pi40	Pi
83	pi60	Pi
84	piec	Dispositif
85	pil	Dispositif
86	plag	Intervalle
87	plausibl	Probabilite
88	poisson	Binomiale
89	popul	Population
90	possibilit	Possible
91	possibl	Possible
92	pourcent	Proportion
93	pourcentag	Proportion
94	precis	Precision
95	probabilit	Probabilite
96	probabilt	Probabilite
97	probabl	Probabilite
98	problem	Probleme
99	proport	Proportion
100	quantif	Quantification
101	quantifi	Quantification
102	repet	Variabilite
103	resultat	Resultat
104	rheto	Cours
105	risqu	Risque

106	sain	Contexte
107	simpl	Simplification
108	simplif	Simplification
109	simplifi	Simplification
110	stabilis	Stabilite
111	stabilise	Stabilite
112	stabl	Stabilite
113	succ	Binomiale
114	syndrom	Contexte
115	tabl	Binomiale
116	taill	Tirage
117	tir	Tirage
118	tirag	Tirage
119	truc	Truc
120	vaccin	Contexte
121	valeur	Valeur
122	variabl	Variable
123	varient	Variabilite
124	victori	Contexte
125	virus	Contexte
126	vraisembl	Probabilite
127	yamagat	Contexte

TABLE 4.3: Association entre les racines et les verbes

	Racine	Verbe
1	analys	Observer
2	compr	Comprendre
3	compren	Comprendre
4	comprend	Comprendre
5	comprendr	Comprendre
6	connaiss	Connaitre
7	connaitr	Connaitre
8	croir	Croire
9	crois	Croire
10	defin	Definir
11	devoir	Devoir
12	devr	Devoir
13	dois	Devoir

14	doit	Devoir
15	faudr	Falloir
16	faut	Falloir
17	imagin	Imaginer
18	imaginon	Imaginer
19	impress	Impression
20	justif	Justifier
21	justifi	Justifier
22	justifie	Justifier
23	observ	Observer
24	obten	Obtenir
25	obtenu	Obtenir
26	obtient	Obtenir
27	peut	Pouvoir
28	peuvent	Pouvoir
29	peux	Pouvoir
30	pouv	Pouvoir
31	pouvoir	Pouvoir
32	sais	Savoir
33	sait	Savoir
34	sav	Savoir
35	savoir	Savoir
36	suff	Suffire
37	suffis	Suffire
38	toler	Tolerer
39	tolere	Tolerer
40	voi	Voir
41	voir	Voir
42	vois	Voir
43	voit	Voir
44	vus	Voir